

P.S. [

Depository

SHL
WITHDRAWN



THE
BRITISH JOURNAL
OF
PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF

A. C. AITKEN	L. S. PENROSE
M. S. BARTLETT	J. FRASER ROBERTS
W. G. EMMETT	A. RODGER
M. G. KENDALL	B. BABINGTON SMITH
D. N. LAWLEY	W. STEPHENSON
P. E. VERNON	

Managing Editor : D. W. HARDING

Volume I
1947-48

UNIVERSITY OF LONDON PRESS LTD.

WARWICK SQUARE, LONDON, E.C.4.



Digitized by the Internet Archive
in 2026

EDITORIAL

The British Psychological Society has decided to issue a new *Journal* dealing with the application of mathematical methods to psychological problems. It will be entitled *The British Journal of Psychology (Statistical Section)*; but the term statistics will be interpreted to cover every form of quantitative approach.

The issue of yet another psychological journal, particularly at this time, calls perhaps for a word or two of explanation and defence. From the days of Galton onwards, British psychology has been foremost in the employment of statistical methods. Yet hitherto it has been exceedingly difficult to print or publish researches using such techniques. Quite rightly editors of the other sections of the Society's journals have hesitated before accepting papers that are largely mathematical, since the vast majority of subscribers not unnaturally protest if they receive an issue filled with articles that are almost wholly unintelligible to them. Hence, writers who may have based their conclusions on statistical analyses have commonly been compelled to abridge their tables and to omit their algebraic proofs, if they wished to secure publication in this country. Psychologists abroad have thus been led to imagine that British psychology, which at the beginning of the century was unquestionably taking the lead in statistical techniques, has during the last few decades made little or no new contribution.

The need for such a journal has indeed long been felt. In this country, until quite recently, readers of psychological articles who are interested in or able to follow mathematical discussions of a somewhat technical nature have been comparatively few. But during the last ten years their number has greatly increased; and there has been a still more rapid increase in the number of investigators with valuable material to publish. Further, the large amount of psychological work carried out for the three Fighting Services during the war has led to the accumulation of extensive and instructive data calling for statistical analysis which will urgently need publication in the near future.

In the new journal it is proposed to print papers dealing both with new methods and with new results. The articles will be of varying length, and, it is hoped, will embrace a wide variety of topics. The methods discussed will include psychophysical methods, scaling methods, studies involving newer statistical techniques (such as analysis of variance and covariance), as well as those based on ordinary correlational procedures or factor analysis. The results will be drawn from the widest possible fields—from individual psychology as well as general psychology, and from the social as well as the educational, vocational, and medical branches. In addition to articles, we propose to include shorter communications in the form of notes, discussions, abstracts, and correspondence, and the usual critical notices and book reviews.

At the outset the papers issued will not be highly technical. There is a growing number of readers who are interested in mathematical approaches and who would like to learn something of their nature and value. Hence many of the articles will be comparatively elementary, and at times, it may be, purely explanatory. We shall also welcome well-informed papers criticizing the quantitative approach, or indicating what the critics take to be its limitations or possible fallacies.

The publication of a separate journal dealing with statistical psychology must not be taken to imply that British psychologists look upon statistical or quantitative psychology as a separate or self-contained department. Mathematics, we have been lately taught, is merely a special branch of logic ; and consequently mathematical techniques must be regarded as merely a special form of general scientific methodology. Our aim, therefore, is not to substitute a "mathematical school of psychology" for the experimental, clinical, or other "schools." In psychology, as in other sciences, the divorce of the mathematical approach from the observational or experimental would be deplorable ; and our hope is, not to encourage a separation between the two, but rather to lead the advocates of alternative approaches to appreciate the fact that a mathematical analysis, *as a supplement* to other methods, may result in a great gain in logical vigour and cogency.

It follows that we shall ask our contributors to cast their papers in a form which shall have as wide an appeal as possible. If in fifty years' time the need for a distinct statistical journal disappears, we shall regard our work, not as a failure, but as a success. Our policy, therefore, will be much like that proposed by Galton for *Biometrika* : "Technical terms and symbols may be necessary ; but it is well to minimize their use. The guiding principle should be to restrict technicalities so far that a competent scientific reader, familiar with kindred branches, should be able to understand the papers without difficulty. Logical arrangement is the first essential."

It is unnecessary to set out in detail any requirements in regard to the form of manuscripts : these have been enumerated more than once in excellent articles issued from time to time by editors of existing psychological journals.¹ In our capacity as co-editors of other periodicals we have been greatly struck with the readiness of psychological contributors to leave to the busy editor such details as the correction of spelling, punctuation, and grammar, the insertion of headings and initial list of contents, and of the pages, dates, and initials of authors in the list of references at the close. We would earnestly appeal to intending contributors to submit their manuscripts, so far as possible, in a shape which can be forwarded at once to the printers, if the article is accepted.

To begin with, numbers will appear, not at regular intervals, but as suitable material accumulates. We hope to issue at least two numbers a year, and later to increase them to three, by which time we hope that regular dates of issue may become practicable. Articles and other material for publication (notes, letters, or book reviews) should be sent to one or other of the editors—either to Sir Cyril Burt, Department of Psychology, University College, London, W.C.1, or to Professor Godfrey Thomson, Moray House, Edinburgh, 8.

C. B.

G. H. T.

¹ Cf. *Psych. Bull.* XXXVI (1939), pp. 25-30.

A COMPARISON OF FACTOR ANALYSIS AND ANALYSIS OF VARIANCE

By CYRIL BURT

Department of Psychology, University College, London

I. *Analysis by Simple Averaging* : (a) *Averaging Successive Deviations* ; (b) *Summation with Equal Weights*. II. *Analysis of Variance* : (a) *With Unstandardized Marks* ; (b) *With Standardized Marks*. III. *Factor Analysis* : (a) *With Simple Summation* ; (b) *With Weighted Summation*. IV. *Final Comparison*. V. *Summary and Conclusions*.

I. ANALYSIS BY SIMPLE AVERAGING

Aims of the Two Methods. The primary aims of factor analysis in psychology are twofold—to classify mental traits and to classify persons. Since the persons will be classified according to their mental traits, the two aims can generally be carried out simultaneously. Further, as psychology is an inductive rather than a deductive science, the actual procedure will consist, not in deducing factors from empirical data, but in confirming hypothetical classifications, already explicitly or implicitly assumed, by applying a statistical test to their corollaries.

Now these are aims which, in principle at any rate, are also achieved by the analysis of variance ; and in several previous publications¹ I have pointed out the close analogy between the two modes of approach. My argument has been that, if, as I believe, both methods are equally valid when applied to psychological data (provided the appropriate conditions are fulfilled), then their results should be fundamentally the same.²

Argument. It is the object of this paper to substantiate and illustrate the comparison suggested a little more fully. In my previous discussions of the problem, I have, as a rule, assumed that the psychological reader would be more familiar with factorial analysis, and consequently endeavoured to describe analysis of variance in factorial terms. In the present paper I shall reverse this procedure.

I shall begin by showing that the plain man's method of totalling or averaging an individual's marks (or their differences) yields a rough but close approximation to what the psychologist would call a 'factor-measurement,' and shall then show that these averages are in fact the measurements of the components or 'sources of variation' which are formally tested in an analysis of variance. In these cases the averages are calculated with equal weights. In the second half of the paper I shall show that factor analysis is concerned with precisely the same components, but seeks to assess them as weighted rather than unweighted (or equally weighted) averages ; and what is called 'factor analysis by weighted summation' is in effect a short cut for determin-

¹ E.g., *Factors of the Mind* (1940), Chap. X, and 'Mental Abilities and Mental Factors,' *Brit. J. Educ. Psych.* (1944), XXIV, pp. 57-68. More detailed illustrations had been given in an earlier L.C.C. Report on 'Classification of Pupils according to Special Abilities,' from which some of the following tables have been taken. I have also to express my thanks both to Dr. M. S. Bartlett and to Dr. D. J. Finney, who have been good enough to read my manuscript in draft and to make helpful and suggestive comments.

² A brief demonstration of this was attempted in an L.C.C. Report on 'The Technique of Marking' ; and Miss Cast, at my suggestion, carried out a more systematic comparison in her thesis on 'Methods of Marking English Composition' (Univ. London Library). In both cases the procedure was to apply analysis of variance and factor analysis to the same data : the essential results are summarized in *Brit. J. Educ. Psych.*, IX (1939), pp. 257-269, and X, pp. 49-60.

ing the best possible weights, i.e., those that yield a least-squares fit. Since in this latter case the weighting matrix is demonstrably orthogonal (like the matrix of equal weights implicitly adopted in analysis of variance), I shall argue that we may test the significance of the factors so determined by the same procedure that is employed in an ordinary analysis of variance. Incidentally, I believe, the detailed comparisons which the discussion will entail will make it easier to recognize the general conditions and the special type of problem for which one mode of analysis rather than the other is to be preferred.

The proofs I shall set out in as simple a fashion as possible, relying rather on arithmetical illustrations than on algebraic deductions.¹ The latter can readily be supplied by the mathematical reader from the few formulæ I shall cite. Hitherto, in expositions of both techniques, though numerical examples are frequently used to explain the working procedures, the underlying theory has been presented almost entirely in abstract symbols ; and hence the full implications of both procedures have been left obscure and even unintelligible to the ordinary psychologist. I hope, therefore, that the more concrete approach that I have adopted may make the aims of both techniques a little clearer, and may prove suggestive to those who are engaged in teaching them. At the same time, the analogies that emerge seem themselves to raise new and unsolved problems worthy of the attention of the statistical expert.

A Practical Problem. My own early efforts in developing a theory of multiple factor analysis were prompted more by a need to classify persons (particularly school children) than to analyse the mind. The former approach is, I fancy, easier for the novice to appreciate ; and has come to the fore once again in recent work for the Fighting Services. Accordingly, in the hope of rendering the argument easier to follow I shall start with a concrete example drawn from the field of education—the field in which factorial analysis (at any rate in this country) has been chiefly developed.

A practical problem which exercised the school psychologist in the early days of his work was to discover the best methods of testing or assessing school children with a view to deciding (i) which are sufficiently bright to be recommended for some kind of higher education, and (ii) which of these brighter pupils are more suitable for a secondary education of the grammar school type and which are more suitable for a trade school or a technical institute. In many areas the initial recommendations were made by the head teacher. For such purposes he would often set his pupils an examination, not unlike the familiar type of scholarship examination. Accordingly, let us take the results of one such examination, comprising (we will suppose) four tests of the four relevant subjects of the elementary curriculum.

The marks shown in Table I were actually obtained at a composite examination of this sort. In order to simplify the tables, I have limited the number of persons to six boys from a single school ; but they have been chosen to represent typical cases and to yield much the same correlations as the entire group.²

¹ In previous contributions I ventured to suggest that analysis of variance could be aptly elucidated if its problems and procedures were re-expressed in the matrix-notation now common in factorial work. The same holds good of the arguments presented below. Nevertheless, as many psychological readers are unfamiliar with the terminology of matrix algebra, I shall here avoid such technicalities so far as possible.

² This particular examination was taken by 252 boys, aged 10 to 11, at seven different schools. But it would have been plainly impossible to illustrate my points with the complete tables for over two hundred boys. The reader, however, must bear in mind that an analysis based on 6 cases only is necessarily artificial. To facilitate my illustrations I have deliberately picked an examination yielding rather high correlations, and have smoothed the marks to give a balanced set of general averages with a standard deviation of 10 ; to provide a basis of comparison for the other pupils I have also included a boy (Hugh) whose marks have been rounded off so that they appear exactly equal to the average in all four tests. This makes the example yet more artificial, since, taken strictly, it would still further reduce the degrees of freedom.

The first was a paper in Problem Arithmetic ; the second required the pupil to write four short English compositions ; the third consisted of exercises in Drawing (‘ geometrical ’ rather than ‘ artistic ’) ; and the last was a practical examination in Handwork, but here the marks were amalgamated with those obtained for manual work throughout the term. Each test was set and marked by a teacher who was in some degree a specialist in the subject. Two days were required for the examination : Arithmetic and Drawing were set on the first day (Monday) ; English and Handwork on the second (Tuesday).

TABLE I. MARKS FOR SIX BOYS IN FOUR TESTS

Tests	Day	Persons						Average
		John	James	Hugh	George	Ralph	Henry	
Academic								
(1) Arithmetic ..	Mon. ..	16	11	8	6	7	0	8
(2) English ..	Tues. ..	15	14	9	5	6	5	9
Manual								
(3) Drawing ..	Mon. ..	17	8	10	16	8	1	10
(4) Handwork ..	Tues. ..	20	11	13	13	15	6	13
Average ..		17	11	10	10	9	3	10

(a) Averaging Successive Deviations

The Common-sense Approach. My first object will be to demonstrate that the procedure adopted by the factorist, like so many techniques of science, is little more than a refinement of the naïve logic of common sense. How then would a practical teacher proceed ?

To determine which is the brightest, the natural course would be to *add each boy's marks* for all four subjects, just as they stand, and compare the totals (or averages) for the different individuals. To determine which of these brighter boys are more suited to a grammar school education as contrasted with a technical education (or vice versa), we may similarly compare each boy's total (or average) for the two academic subjects with his total (or average) for the two manual subjects.

This, of course, assumes that the marks in each test are awarded by examiners on a comparable basis. Some degree of equivalence is commonly ensured by prescribing the same maximum for each subject—here 20. However, from the average marks obtained in each test by the group as a whole it certainly looks as though some of the tests (or possibly some of the examiners) were more severe than others ; and it was further suggested that most of the candidates seemed to work better on the second day than on the first. The headmaster supervising the examination therefore proposed to eliminate these further differences as irrelevant, and accordingly reduced the marks sent in by each examiner to terms of the same average, namely, 10. His reduction is shown in Table II.

TABLE II. MARKS IN THE FOUR TESTS REDUCED TO THE SAME AVERAGE

Test	John	James	Hugh	George	Ralph	Henry	Average
(1) Arithmetic	18	13	10	8	9	2	10
(2) English	16	15	10	6	7	6	10
(3) Drawing	17	8	10	16	8	1	10
(4) Handwork	17	8	10	10	12	3	10
General Average ..	17	11	10	10	9	3	10
I General Average as a Deviation	+7	+1	0	0	-1	-7	0
II Average Difference ..	0	+3	0	-3	-1	+1	0

The 'general average,' given at the foot of the table, shows the order of merit of the six boys for 'general ability'; the 'average difference,' i.e., the difference between a boy's general average for all four subjects and his average for the two academic subjects, shows how far he is superior in grammar school subjects as contrasted with manual, or vice versa, i.e., his special aptitude. *The last two rows of figures, then, give straight away a plausible solution to our two main questions:* we see that John is the brightest of the six, and that James has a special aptitude for academic subjects and George for practical. The rows therefore (marked I and II respectively) may be regarded as our first estimates for the two main factors measured by these tests. But now let us note that this method of calculating the measurements for the two types of ability *does not guarantee that they will be absolutely independent*, as we should expect them to be from the way those abilities are defined. In Table II, for example, the correlation between the last two lines is 0.067. Thus this simple procedure leaves our estimates of a boy's *relative* ability in the two types of subject still somewhat dependent on our estimate of his *general* ability, whereas that should of course be ruled out altogether when we turn to assess special aptitudes.

Moreover, there would seem to be further influences affecting the marks— influences which are irrelevant to our main issue, and which yet, if appreciable, ought not to be altogether ignored. For instance, it is conceivable that, besides the general effect on the class as a whole, certain individual boys may be more seriously handicapped by taking certain tests on the first day rather than the second (or vice versa): and these individual differences will not be eliminated merely by deducting an allowance for certain days from the marks obtained by all the candidates. To complete our study, therefore, we ought to average each boy's deviation for the first day and for the second, and then compare the two sets for every individual.

(b) Summation with Equal Weights

An Equivalent Mode of Calculation. Now, a little algebra will quickly show that instead of averaging all four marks, and then averaging deviations from these averages, and finally averaging these averaged deviations, *we can reach precisely the same figures by a simple process of weighted summation*, using weights which differ solely in the distribution of their plus and minus signs and have *the same numerical value* throughout, viz., the reciprocal of the number¹ of tests employed (i.e., here $\frac{1}{4}$). The scheme of weighting is shown in the left-hand portion of rows 5 to 8 in Table III.

¹ As will appear later, it would be slightly more convenient to divide throughout by $\sqrt{n}$ instead of n (provided n has a suitable value: see footnote 2, p. 8). Here, therefore, the factor weights would be $\frac{1}{4}$ instead of $\frac{1}{2}$. But to begin with I have taken the ordinary method of averaging, since that would seem most natural to the ordinary reader.

TABLE III. OBSERVED MARKS IN DEVIATION FORM: ANALYSIS AND RECONSTRUCTION BY EQUAL WEIGHTS

Test				John	James	Hugh	George	Ralph	Henry	Sum of Squares		
(1) Arithmetic	..	..	..	+8	+3	0	-2	-1	-8	142		
(2) English	..	..	..	+6	+5	0	-4	-3	-4	102		
(3) Drawing	..	..	..	+7	-2	0	+6	-2	-9	174		
(4) Handwork	..	..	..	+7	-2	0	0	+2	-7	106		
Total Square ..										524		
Test Weights				Analysis into Factor-measurements								
	(1)	(2)	(3)	(4)								
I	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	..	+7	+1	0	0	-1	-7	100
II	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	..	0	+3	0	-3	-1	+1	20
IV	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	..	$+\frac{1}{2}$	$-\frac{1}{2}$	0	+2	$-\frac{1}{2}$	$-\frac{1}{2}$	7
III	$\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	..	$+\frac{1}{2}$	$-\frac{1}{2}$	0	-1	$+\frac{1}{2}$	$-\frac{1}{2}$	4
Total Square ..										131		
Factor Weights				Reconstruction of Test-measurements								
	I	II	IV	III								
(1)	1	1	1	1	..	+8	+3	0	-2	-1	-8	142
(2)	1	1	-1	-1	..	+6	+5	0	-4	-3	-4	102
(3)	1	-1	1	-1	..	+7	-2	0	+6	-2	-9	174
(4)	1	-1	-1	1	..	+7	-2	0	0	+2	-7	106
Total Square ..										524		

The four weights (all positive) forming the first line on the left (row 5) are used to multiply the corresponding four rows of marks shown above in the body of the table (rows 1 to 4); and the weighted figures are then summed: the result, of course, is the same as that of adding each column of marks as they stand, and then dividing by 4, to get each boy's average. In much the same way, the second row of weights gives us, by a shortened procedure, the averaged difference for the two types of subject; and the third row gives us the averaged differences for the two days. If now we assume that the plus and minus signs must remain equally balanced, only one other type of weighting is possible, namely, that given by the last row of weights. With this, as the pattern of signs plainly indicates, the figures obtained will express the average difference between the averages (*a*) for Arithmetic and Handwork, and (*b*) for English and Drawing. Accordingly the psychologist might be tempted to speculate whether these last results may not indicate some further aptitude or capacity—e.g., some specialized ability common to the first two subjects (such as facility in quantitative measurement and computation) or to the second two (such as æsthetic self-expression). But, as a little algebra (or an actual trial) will show, these last figures represent the averages of the final residuals left after deducting the other three weighted averages from the original marks. Consequently, we may provisionally regard them as measuring the amount of uncontrolled variation in the differences that are due chiefly to the three main controllable causes. With this interpretation we can employ this final row to get a *first rough estimate of the errors or irrelevances entailed* in taking these particular test-performances as samples of what we chiefly want to measure.

The weighted sums obtained by these four modes of weighting are shown in rows 5 to 8 under the names of the six boys. And these figures may accordingly be taken as measuring for each boy (i) his general ability, (ii) his special aptitude for

abstract subjects as compared with manual, (iii) his liability to be affected more favourably by the first day of the examination as compared with the second, and (iv) the irrelevant variations to which he is subject in particular tests and on particular days.¹ From these four rows of hypothetical measurements we can reconstruct with perfect accuracy the four-rowed set of marks from which we started. We can therefore consider these four components as additive 'factors,' which will completely account for the observed test-marks, and regard each row of figures as expressing the pupils' 'factor-measurements.' To effect the reconstruction we need a *second matrix of weights which will operate as the reciprocal* of the first. This second set of weights is shown beneath the first; and the reader can verify the reconstruction for himself.

Anticipating their reappearance in a more elaborate shape, I shall ask the reader to observe two characteristics of these weighting matrices: in each (i) the sum of every row (except the first) is zero, and (ii) the sum of the cross-product of any row by any other row is always zero, i.e., each row is uncorrelated with the rest.²

II. THE ANALYSIS OF VARIANCE

The Theoretical Problem. What I have called the common-sense procedure, or something very like it, has been in practical use for a generation or more.³ But few have attempted to examine the validity of the results that it provides. Accordingly, the first task of the theoretical psychologist, who takes up this field of work, is to discover what justification there may be for such a procedure, and then, by the aid of this more explicit knowledge, to correct or refine it.

First of all, can we be sure that the final figures we are proposing to use for promoting and classifying these pupils are determined by anything more than mere random fluctuation? The practical bearing of the question is clear. Under the new scheme for post-primary education, for instance, our headmaster (looking at rows 5 and 6 of Table III or the last two rows of Table II) would probably decide that James should go to a 'grammar' school, George to a 'technical' school, Hugh, Ralph, and Henry to a 'modern' school, and that John would shine in any type of school. But, before discriminating between Ralph and George in this way, he ought first to make sure that Ralph's 'general average' and his 'average difference' differ significantly from those obtained by George.

¹ In the table the four 'factors' have been numbered in the order in which they emerge in the *final* method of analysis (viz., with 'weighted summation,' see below, p. 23).

² It will be noted that the effect of such 'weighting matrices' is not so much to *weight* the tests as to *classify* them, and to classify them dichotomously. They thus play an important part in the theory of the transformation of both traits and group factors into bipolar factors, and vice versa; and may therefore be called 'transclassificatory operators.' Numerically they have the form of 'equilibrated' orthogonal matrices: for, in general, the weighting elements are only equal when the number of tests or traits is some integral power of 2 (i.e., 2, 4, 8, 16, . . .). In other cases they are equally balanced rather than equal. In the commonest type every row except the first has only $m+1$ elements (where m is the number of the row in inverse order); and the elements have the values $(1, m, \text{ or } m+1) / \sqrt{m(m+1)}$. This type is therefore 'triangular'; (cf. *Marks of Examiners*, 1936, p. 307, *Psychometrika*, III, 1938, p. 160, and references). Since each row of positive and negative weights must add up to zero, the more general form of the procedure described in this section might be termed 'Summation with Equal Balanced Weights.' From this it will be seen that we ought to aim at a similar balance in selecting our tests.

³ Thus at my own school the headmaster used to set an annual examination for boys at the top of the Lower School; and, on the basis of the results, would first pick out those who were to be promoted to the Upper School, and then divide them into two main groups—those going to the Classical side and those going to the Modern or Mathematical side respectively. For this purpose he adopted a double process of averaging and taking differences almost identical with that described in the text.

He might begin by comparing the difference between the two boys' marks with its standard error. But since he and his colleagues who helped to plan this investigation were interested, not merely in the fate of this boy or that, but in the validity of the inferences drawn from the examination as a whole, he would not be content merely to test a few crucial or borderline cases, but would want to confirm every conceivable difference. And, if he is up to date in his knowledge of statistical techniques, he will know that all the differences can be tested simultaneously by undertaking what is known as an analysis of the variance.

(a) With Unstandardized Marks

A Hypothetical Three-way Classification. Now the figures just calculated lend themselves directly to this mode of testing. Indeed, the method I have described for analysing the observed marks into four independent components by *simple averaging yields precisely the same results as those that would be obtained on analysing the variance* in the usual way. The student should test this for himself by calculating the variances according to the usual routine.¹ Here, however, it will be more instructive to make a complete analysis of the initial data. If we begin with the original table of marks (Table I), we shall then be able to test, not merely the significance of the particular influences in which we are chiefly interested (the boys' abilities and special aptitudes), but also the significance of those irrelevant factors we wish to eliminate (e.g., the difficulty of the tests and the effect of the day on which they were taken).

In theory we can separate such data into as many factors as there are degrees of freedom; but most of the factors would possess neither statistical nor psychological significance. In analysis of variance, therefore, the hypothetical classification with which we can most profitably start is usually based on our prior knowledge of the variables concerned: indeed, it is this prior knowledge which determines beforehand our selection of variables to be measured. Here it is natural to think first of the possible differences in the general ability of the several boys and in the general difficulty of the several tests. But it is conceivable that the tests may also differ according to the type of subject involved (whether academic or manual), and according to the day on which they were given; and the individual boys may differ according to their aptitude for this or that type of subject, and according to their condition on this or that particular day—a mode of variation commonly named 'interaction.'

The reader, familiar with factorial procedures, will note that the 'interactions' are interactions between principles of classification for persons (columns) and for rows (tests). Those of the first order correspond with the bipolar factors which, we shall find, are ordinarily obtained by factorizing the correlations between the tests. The fact that the interactions are mutual, however, implies that we should reach the same bipolar factors if we started by correlating persons. Here, therefore, we have a further confirmation of the result which I have sought to demonstrate elsewhere, viz., that, with all except the first or general factors, there is a reciprocity between the factors obtained by correlating tests and the factors obtained by correlating persons.

The general mode of classification which I have just described suggests testing *three main effects* to begin with—those due to differences in persons, type of subject, and day. These three main sources of variation will entail three 'interactions' of the first order, and one 'interaction' of the second order. As usual, the last (being the interaction of highest order) will be used to furnish our estimate of sampling

¹ See, for example, Snedecor, G. W., *Statistical Methods*, pp. 216 et seq.

error. With this scheme of classification the requisite square-sums,¹ mean squares, and variance ratios can be calculated in the customary way. The number of boys in our sample is $N=6$. But, since the figures in each row of the components are deviations about the mean of zero, each square-sum is based on only $(6-1)=5$ degrees of freedom. Accordingly, to obtain the mean square or 'variance,' we must divide the square-sums by 5. The final analysis of variance can therefore be tabulated in the usual way as follows (Table IV).

TABLE IV. ANALYSIS OF VARIANCE: (a) WITH UNSTANDARDIZED MARKS

Factor	Source of Variation		Degrees of Freedom	Sum of Squares	Mean Square	Ratio of Variances	Probability
I	..	A. Persons	5	400	80	25.00	0.001-0.01
—	..	B. Type of Subject..	1	54	54	16.87	0.01
—	..	C. Day	1	24	24	7.50	0.01-0.05
II	..	AB. Persons and Type of Subject ..	5	80	16	5.00	0.05-0.20
III	..	AC. Persons and Day ..	5	28	5.6	1.75	<0.20
—	..	BC. Type of Subject and Day ..	1	6	6	1.87	<0.20
IV	..	ABC. Residual ..	5	16	3.2	—	—
Total ..			23	608	—	—	—

The Test of Significance. We shall begin then by taking the last set of variations, i.e., what we have called the 'residual' variance, as supplying our initial estimate of error, and proceed to divide each of the other three variances by this estimate to obtain the critical ratio. On referring to the usual probability tables² for variance-ratios, we at once find that, with this artificially small sample, none of the interactions is significant. Had this result been obtained from the complete sample, we might then have gone on to pool³ the interactions, in the hope of reaching a more reliable estimate of error based on a larger number of degrees of freedom. With the figures given in the table we should conclude that the influence of the different days is also

¹ On the analogy of the familiar term 'product-sum,' I have found it convenient to coin the phrase 'square-sum.' 'Covariance' and 'variance' are now generally used to denote the mean 'product-sum' and the mean 'square-sum' respectively, the means being usually estimated by dividing by the degrees of freedom rather than by the precise number of products or squares that have been summed. When deviations are in *unitary* standard measure, the terms 'covariance' and 'variance' are sometimes used for the product-sum and square-sum of the figures taken as they stand (i.e., without any further division), since they coincide with the 'covariance' and 'variance' as calculated by division from *ordinary* standard measure: provided the context makes the meaning clear, this usage may be convenient, if not entirely consistent. However, it is frequently handy to have brief technical terms for what is sometimes called the 'unaveraged covariance' (or 'variance'); and for this purpose the terms 'co-deviance' (and 'deviance') may perhaps be suggested.

² Fisher and Yates, *Statistical Tables for Biological, Agricultural, and Medical Research* (1943), Table V, pp. 32-39. With this small sample, of course, no real importance can be attached to the final column of probabilities. They are included merely to illustrate the general mode of tabulation.

³ This method of partitioning the data could be defended on the ground that it gives an indication (somewhat artificial, no doubt) of the effects of the 'specific factors' involved in each of the four subjects by pairing them in the two remaining ways. The reader, who thinks of tests as dependent chiefly on 'abilities,' may have supposed that this was the more natural scheme to have adopted at the outset: but, since there was some *a priori* likelihood that the day of the test might also have a significant influence, it seemed more instructive to examine this mode of approach first.

non-significant. Proceeding in this way, and pooling¹ all the non-significant variances, we should eventually discover that a low significance attaches to the differences in difficulty between the four tests, and a comparatively high significance to the differences between the persons tested.

Now neither the difficulty of the tests nor the effect of taking them on different days is of special interest to us here. Accordingly, we can rule out any possible effects of this kind by working, as before, with *deviations about the averages for the several tests*: that at once eliminates from our analysis (B) type of subject, (C) day, and (BC) the interaction between the two. If we imagine the figures thus designated deleted from Table IV, we are left with the four rows labelled "Factors I, II, III, IV" respectively.² With this 'reduced sample' of data the total sum of squares is now reduced to 524. This was, in fact, the 'total square' analysed in Table III. The four components are also the same;³ and the probabilities deduced from the variance-ratios remain unaltered.

(b) With Standardized Marks

So far, however, we have treated the variances of the several tests as if they and their differences possessed objective validity. But this is extremely doubtful. For example, on looking back at the original marks (Table I) we see that the range of marks for Arithmetic and Drawing was 16, whereas for English it was only 10. To prove that children do in fact vary more widely in Drawing or Arithmetic than in English composition would be far from easy; and it seems more likely that such differences are merely the accidental result of the way each examiner has marked the test-papers. Accordingly, in the absence of more adequate information, a safer plan will be to adopt *precisely the same type of scale for every test*. This can be done by reducing the original marks to *unitary standard measure*, i.e., to a series with the same standard deviation (1.0) as well as the same mean (0.0). The marks for the four tests rescaled in this way are shown in Table V.

These basic figures are given to four decimal places, to enable the reader to check the subsequent computations (the actual working has been carried to six decimal

¹ Whether such pooling is justifiable must not, of course, be decided solely by the fact that the items in the table under consideration happen to be devoid of statistical significance, as judged by the common convention. 'Non-significance' does not imply that a conceivable effect has been disproved, but merely that the evidence available is not sufficient to prove it. Thus, when its mean square is larger than the residual mean square, the presence of a specific effect may still appear to be more probable than not, even when its probability is too low to be judged technically 'significant.' The issue must be settled partly by antecedent or external evidence, as well as by the figures actually obtained, i.e., by the prior as well as by the posterior probability in its favour. On this question the reader may usefully consult Gouliden, C. M., *Methods of Statistical Analysis* (1939), ch. XI, sect. 6.

² These four rows correspond, as we shall discover later, to the four factors derived from 'correlating tests': the rows rejected correspond with the factors we should derive by 'correlating persons'.

³ Since each of these four components appears four times over in the initial table—once in each test—we have to multiply the component squares as given in Table III (viz., 100, 20, 7, and 4) by 4 to obtain the square-sums left in the reduced form of Table IV. The need for this supplementary step is obvious, when we compare the sum of the squares for each of the components as given in Table III with the sum of the square-sums for the rows of original (or reconstructed) marks: these are given in the last column of that table; and the sums are seen to be 131 and 524 ($=4 \times 131$) respectively. The two steps could be condensed into one by taking for the weights, not $1/n - \frac{1}{4}$, but $1/\sqrt{n} = \frac{1}{2}$ (as suggested above). It will then be seen that the complete weighting matrix required for the analysis of variance is a *simple orthogonal matrix whose elements consist of the reciprocal of the square root of the number of traits*; and the matrix required for reconstructing the marks will be the *inverse or reciprocal* of this matrix, which is merely the same matrix transposed, since the transformations are orthogonal.

places). In theoretical studies the psychologist nearly always prefers to work with marks re-standardized in some such fashion, though for practical purposes the figures may seem needlessly cumbersome. Accordingly, let us briefly glance at the more precise results secured by averaging these re-standardized figures¹ along the same lines as before.

TABLE V. MARKS REDUCED TO UNITARY STANDARD MEASURE

Test	John	James	Hugh	George	Ralph	Henry
(1) Arithmetic	·6713	·2517	·0000	—·1678	·0839	—·6713
(2) English	·5941	·4951	·0000	—·3961	—·2970	—·3961
(3) Drawing	·5307	—·1516	·0000	·4549	—·1516	—·6823
(4) Handwork	·6799	—·1943	·0000	·0000	·1943	—·6799
Average	·6190	·1002	·0000	—·0273	—·0846	—·6074
Average in U.S.M. ..	·7054	·1142	·0000	—·0310	—·0964	—·6922

The Modified Factor-measurements obtained by Averaging the Marks in Standard Measure. The averages for each boy are given in the last line but one of Table V. The last line of all reduces these factor-measurements themselves to unitary standard measure; and if we multiply its decimal fractions by 10, we shall have a set of averages not very different from those obtained in Table III (line 5).

It is unnecessary to repeat all the calculations in detail: these the reader, if he wishes, can make for himself. We shall find that the factor-measurements reached for each boy by simple averaging are not very greatly changed; but the changes, slight as they are, tend to become appreciably larger, as we pass from the first factor to the fourth. Table VI gives an analysis of variance for the reduced sample of data,

TABLE VI. ANALYSIS OF VARIANCE: (b) WITH STANDARDIZED MARKS

Factor	Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Ratio of Variances	Probability
I ..	A. Persons	4	3·0810	·7702	21·94	0·01-0·001
II ..	AB. Persons and Type of Subject	4	·6257	·1564	4·46	0·05-0·20
IV ..	AC. Persons and Day ..	4	·1521	·0380	1·08	0·20
III ..	ABC. Residual	4	·1403	·0351	—	—
	Total	16	4·0000	—	—	—

¹ This method of obtaining the 'general factor' (as we should now call it)—namely, reducing the observed measurements to terms of the same statistical unit (e.g., the quartile deviation) and then taking the simple sum—was first suggested by Galton. And he proposed to assess the importance of the factor by calculating the "error" of such sums: "it would be $\sqrt{n}$, if all the variables were perfectly independent; n if they were perfectly related. The observed value is almost always somewhere between these extremes, and would give the information wanted" (*Proc. Roy. Soc.*, XLV, p. 145). The "observed value" to which Galton refers is the root of the sum of the $n \times n$ inter-correlation table, with the self-correlations taken as unity (i.e., $\sqrt{\sum \sum r_{ij}}$).

based on the standardized marks.¹ The probabilities and general conclusions remain as before.

III. FACTOR ANALYSIS

A Factorial Interpretation. It seems clear that the simple procedures I have so far described can readily be expressed in terms of factor analysis, provided we are willing to assume (as I have urged elsewhere) that *factors can be regarded as essentially derived by averaging*, i.e., that the general factor derived by correlating traits is merely an average (or sum) of the measurements obtained for each person, and that the bipolar factors subsequently derived can similarly be obtained by averaging the differences which measure the so-called 'interactions.'

Here I shall confine myself to the analogies observable when we correlate tests. But the same principles would hold good of the factors reached by correlating persons. Our primary aim will be to discover what justification there is for averaging these marks as we have done ; and, so far as our two main practical problems are concerned, the factorial argument will run as follows.

Correlating the Tests. The first step in the process (namely, using the grand averages as they stand to assess the most conspicuous differences between the persons tested) would be justified theoretically, if we could show that there is a legitimate ground for postulating some kind of general ability—some common factor which has operated as the chief determiner (though not, of course, the sole determiner) of the pupils' performances in all the tests we have used. In that case each of the four tests could be treated as furnishing more or less imperfect measures of that common ability² ; and the marks obtained by any one boy would be more or less proportional for all four subjects, though of course by no means identical. The averages of the four sets of imperfect measurements could accordingly be accepted as yielding more accurate estimates than any one test taken alone.

If, then, all four tests have some one component in common—'general ability,' as we have called it—*each test should exhibit a positive correlation with every other*. If, on the other hand, we were actually dealing with a set of independent and uncorrelated traits, then the average for all the tests would necessarily tend towards the same general value for every person tested : for in the long run the average measurement for each pupil would (apart from sampling errors) be identical with the average measurement for the whole group. An examination of the correlations should

¹ The effect of fixing the standard deviation as well as the mean involves problems that require further discussion. An ordinary analysis of variance involves the partition of a sum of squares into a number of components defined by linear restrictions. Does the conversion of marks to standard measure infringe this condition ? And how does the theory apply to those cases in which the very scheme of marking yields test-scores which from the first are supposed to be in standard measure. Dr. Bartlett replies (in a written comment he has been good enough to send me) : " Until I see evidence to the contrary I personally would neglect the effect of fixing the standard deviation, and allocate $n-1$ degrees of freedom (i.e., 5 in this example) for a straightforward analysis, and (with four tests) a total of $5 \times 4 = 20$." But in actual practice neither the standardization nor the proposed reduction of degrees of freedom by one is likely to make any serious change in the inferences drawn regarding statistical significance.

² This argument may appear to introduce a change of standpoint. If, following Pearson, we regard what is now called factor analysis as a special case of the general problem of multiple regression (the special condition being that the criterion is now based on internal rather than on external evidence), then the several tests would have the status of *separate* variables ; and the problem is thus envisaged as one of multivariate analysis. But when we have standardized our test-measurements above, and still more when we have combined them by a weighting equation into a set of factor-measurements for a single 'general factor' (e.g., intelligence), we are implicitly treating them as fallible determinations of a *single* variable. Evidently, as so often occurs in statistical theory, the same general problem is capable of quite different approaches. Nevertheless, I believe the results eventually reached from either standpoint will be found consistent in all essential points.

therefore give the same answer as was obtained by testing the significance of the differences between the averages.¹

An analogous type of proof would serve to justify the second stage in the head-master's procedure, namely, the process of averaging the deviations about each boy's average; for if these deviations were uncorrelated, then their averages also would show only minor variations from pupil to pupil attributable solely to chance.

Let us therefore calculate the correlations. Since the test-measurements have all been reduced to unitary standard measure, their simple product-sums will be identical with the product-moment correlations.

TABLE VII. CORRELATIONS BETWEEN TESTS AND CALCULATION OF FIRST FACTOR SATURATIONS BY SIMPLE SUMMATION

Test	Arithmetic	English	Drawing	Handwork	Total
(1) Arithmetic	1.0000	.8808	.7125	.8477	3.4410
(2) English	.8808	1.0000	.3753	.5193	2.7754
(3) Drawing	.7125	.3753	1.0000	.8247	2.9125
(4) Handwork	.8477	.5193	.8247	1.0000	3.1917
Total	3.4410	2.7754	2.9125	3.1917	12.3206
Saturation	.9803	.7907	.8298	.9093	3.5101

Note.—Saturation = Column Total $\div$ ($\sqrt{12.306}$, i.e., 3.5101).

The correlations between tests are given in Table VII. It will at once be seen that all are positive. However, with the small sample we have taken for illustrative purposes, only the largest would be judged significant by Fisher's criterion; nevertheless, the others are certainly high enough to raise a strong presumption that their joint effect may be significant. Clearly, then, it would be convenient if we could devise some method for reducing *each column of correlations to a single figure*, and thus assess the correlational tendency of each test by examining the size and statistical significance of that single figure. As I argued long ago in my first paper on the subject,² the obvious function to take is the sum (possibly the weighted sum) of the correlations for each test. When divided by the standard deviation of the sums of the marks, such a sum at once yields a convenient measure of the hypothetical correlation of each test with the factor, and thus expresses what has been called the 'factor saturation' or 'factor loading' of the test.

The Need for Differential Weighting. At this point a crucial question at once suggests itself. With the procedures so far employed, we have been content with unweighted sums or simple averages; and have thus *tacitly assumed that each of the four tests provided an equally trustworthy clue* to the true measurements for the various factors or components. But the validity of this implicit assumption may itself be questioned.

Consider the first or 'general' factor by way of an example. We can readily calculate the hypothetical correlations between each of the four tests and the provisional measurements for this factor by estimating the several factor-measurements as unweighted sums or averages of the standardized marks. On computing the

¹ In the field of psychological testing this argument was first clearly formulated by Wissler ('The Correlation of Mental and Physical Tests,' *Psych. Rev. Mon. Supp.*, III, 1901, p. 3). The general principle, of course, had been stated by Galton and Pearson, but with physical rather than mental traits chiefly in mind). The arguments which Spearman proposed to substitute in his first article (*Am. J. Psych.*, XV, 1904, pp. 201f.) were proofs, not of the presence of 'general intelligence,' but of its identity with 'general sensory discrimination' and of the absence of any other factor.

² *Brit. J. Psych.*, III (1909), pp. 160-163.

product-sum correlations in the usual way with the figures given in Table V, we obtain the following values (as we shall see in a moment, they are necessarily the same as the ' saturations ' calculated from Table VII) : Arithmetic, .9803 ; English, .7907 ; Drawing, .8298 ; Handwork, .9093. Thus the several tests differ appreciably in the extent to which each correlates with the general factor thus assessed, though with that factor the differences are not large. With the supplementary factors, as we shall see, the correlations tend to differ far more widely, some being not far from zero (cf. Table VIII).

If, therefore, we want the most accurate assessments obtainable for the first or general factor, we manifestly ought not to assign tests like English and Drawing the same weights as we give to Arithmetic or Handwork. Let us then *weight each test differentially, in accordance with its factor-correlation*, and recalculate the factor-measurements as weighted averages or sums. But altering the factor-measurements will necessarily produce alterations in the factor-correlations. Accordingly, we shall have to recalculate the weighted sums afresh with these new correlations as weights. And this will change the factor-correlations still further, though not so much as before. Eventually, after repeating the cycle of calculations again and again, we shall *by successive approximation reach a stable set of factor-measurements* whose correlations with each test are identical with those assumed as weights.

This final set we may regard as the *best obtainable estimates* for the general factor assumed to underlie each of our four tests. There is a convincing reason for regarding these estimates as the best : if we wanted to give each boy a single mark from which we could forecast his performances in such examinations with a minimum of error, this set of measurements would be found to supply the most accurate predictions. They would be most accurate in the sense that the sum of the squares of their divergences from the actual marks would be smaller than those furnished by any other estimates.

The figures thus reached will be found in the top line of Table XIII. So far as the first factor is concerned, although the weight for Arithmetic has been raised to a quarter as much again as the weight for English, nevertheless, as will be seen, the differential weighting has introduced comparatively little change. The changes in the supplementary factors we shall examine later on.

The Need for a Short Cut. Now, in any actual inquiry our mark sheet would be much larger than that shown in Table I. For example, in the investigation from which this illustration is taken, in order to study the distribution-curve of the factor-measurements, marks from the same tests were analysed from over 250 pupils. In such a case, the process of successive approximation, if applied direct to the full table of marks, would prove extremely slow and laborious. Fortunately a little algebra enables us to devise a convenient short cut.

First, it is easy to show that we can calculate the correlation between each test and the factor-measurements direct from the intercorrelations of the tests, without troubling to calculate explicitly the factor-measurements themselves. Secondly, and for much the same reasons, we can apply the process of successive approximation *to the table of intercorrelations instead of to the test-measurements*. This means that, instead of working with the initial $n \times N$ table of marks, we shall work with the smaller $n \times n$ table derived from it by multiplying the table by its transpose, i.e., we shall work with its ' false square '—a common expedient in matrix algebra. The $n \times n$ product-matrix so obtained is, of course, the table of ' unaveraged covariances ' or product-sums, that is (if the marks have already been standardized), the table of correlations with unity in the leading diagonal.

To show how the results achieved by these alternative procedures differ in no

fundamental way from those already reached, we may consider in slightly greater detail the two commoner methods for factorizing such data, namely, those termed 'simple summation' and 'weighted summation' respectively.

(a) The Method of Simple Summation

The General Factor as Obtained by Simple Summation. As we have already seen, to find the correlation of each test with the general factor-measurements, provisionally estimated by the simple sums or averages of the standardized test-marks, we have merely to *add the columns of correlations* (instead of adding the test-marks), and then divide the total for each by the square root¹ of the grand total for the entire table. This simple summational procedure gives us at once, by a quick short cut, the 'saturation coefficients' for the first factor: the figures are appended at the foot of Table VII. By turning to the averages for each boy shown in the last line of Table V, and calculating their correlations with the test-measurements by the ordinary method (as suggested above), the reader can readily satisfy himself that each of the saturation coefficients, deduced by simple summation from the test-correlations, *is in fact identical with the correlation* between the test in question and the general factor, when that correlation is calculated by the more lengthy procedure from the individual measurements.

Supplementary Factors. To determine the next factor, the natural method would be once again to begin with the detailed table of standardized marks, and deduct the effects attributable to the first or general factor. (The contribution of that factor to a particular test can be calculated in the usual way by multiplying the factor-measurements by the regression of the test on the factor, that is, as we shall see in a moment, by the saturation itself. The details of the procedure will be illustrated when we come to the method of weighted summation.) If the residuals left after the first factor has been deducted appear statistically significant (or if their inter-correlations appear to warrant it), we can then use these residuals (or their inter-correlations) to determine a second factor. Now, however, we shall be concerned with the *differences* between the pupils' residual marks in one group of subjects as contrasted with another. Hence we must subtract the residuals furnished by one group (e.g., the technical) from those furnished by the other group (e.g., the academic: the assignment of tests to one group or the other will be decided by their positive and negative intercorrelations). But by the familiar 'rule for parentheses' we can effect the subtraction of the first group from the second by reversing the signs for the tests bracketed to form the subtrahend group, and then summing the residuals algebraically.²

¹ If x_{ij} denotes j 's marks in the i th test, expressed in unitary standard measure, his total mark in all the n tests will be $\sum_i x_{ij}$. The square-sum for all such totals will therefore be $\sum_j \left(\sum_i x_{ij} \right)^2$. Multiplying

out and summing, we find that this is identical with the sum of all the n^2 correlations in the correlation table (the self-correlation being taken as unity). Hence to calculate the factor saturation as above defined for test i , we have merely to divide the sum of the correlations for that test by the square root of the sum of all the correlations; (there is a convenient check on the arithmetic: the square root is exactly equal to the sum of the factor saturations). This sum, indeed, was the quantity proposed by Galton to measure what Pearson termed the "total correlativity of the system." The argument, of course, only holds good of the particular method of factorization above described.

² If the reader takes the residuals obtained when the general factor has been estimated by simple summation, and adds them just as they stand, he will find that the sum for each boy (like the sum of any set of deviations from a mean) necessarily comes to zero. Hence, we can only obtain a non-zero sum, if certain of the signs are reversed. But it would be wrong to excuse this reversal of signs (as is generally done) on the ground that it is a convenient but arbitrary device for avoiding the vanishing of the residuals when summed as they stand. The sign-reversal is in effect a mode of weighting the tests; and, as we have seen in discussing Table III, there is a logical reason for the changes: this will be obvious when the reader has studied Table XII below.

In this way the working procedure will still be 'simple summation.' Having eventually obtained the detailed factor-measurements for the second factor, we could then calculate the factor saturations for each test by correlating this new set of factor-measurements with the several test-measurements as before.

We could repeat the same process to determine a third factor and then a fourth. Since we started with only four tests, all the residuals that remain after four sets of factor-measurements have been subtracted in succession, will necessarily be zero. And generally, with only n tests, n factors at most can be extracted.

But here again it is easy to show that identical results can be reached if we adopt the same short cut as before. Thus, instead of subtracting the measurements for the first factor (duly weighted by the corresponding factor saturations) from the observed test-measurements, and then adding the residuals (with signs appropriately reversed), we can *subtract the products of the saturations themselves from the observed correlations*, and then add the residual correlations (with signs similarly reversed). And, as before, the column-totals must be divided by the grand total of the whole table. Thus the same 'simple summation' formula can be used for the supplementary factors as well as for the general factor. The complete set of factor saturations¹ obtained in this way is shown in Table VIII.

TABLE VIII. FACTOR SATURATIONS OBTAINED BY SIMPLE SUMMATION

Factor					I	II	III	IV	Sum of Squares
(1) Arithmetic	..	..	..	..	·9803	·1829	·0548	·0501	1·0000
(2) English	..	..	..	..	·7907	·6077	—·0548	—·0501	1·0000
(3) Drawing	..	..	..	..	·8298	—·4828	—·2755	·0501	1·0000
(4) Handwork	..	..	..	..	·9093	—·3078	·2755	—·0501	1·0000
Sum of Squares	..	..	..	..	3·1016	·7306	·1578	·0100	4·0000
Contribution to Variance (as a percentage)	..	..	..	..	77·54	18·26	3·94	0·26	

Regressions of Tests on Factors. The factor saturations for a given test do not merely express its correlations with the several factors: they also specify the weights (or 'partial regression coefficients') by which each row of factor-measurements must be multiplied, if we wish to reconstruct the observed test-measurements from the hypothetical factors. They have thus a twofold interest. They indicate how far each of the hypothetical factors enters into and determines the pupils' actual test-performances; and they also enable us to calculate the completeness and exactitude with which we can account for the observed test-measurements by the aid of one or more of the hypothetical factors. If we invoke *all* the factors, we should be able to reconstruct the test-measurements exactly. But, before we can proceed to this check, we must first decide how to calculate the several factor-measurements themselves.

Regressions of Factors on Tests. For the various calculations mentioned, we need, not only the regressions of *the tests on the factors*, but the regressions of *the factors on the tests*. Here the problem is the same as that which arises whenever

¹ Owing to the equality of the weights for the first factor, there are instructive relations between the mean squares as calculated for the analysis of variance and the factor saturations as calculated by simple summation in this way (cf. *Factors of the Mind*, p. 277). As can be seen from Table VI above, with the first two factors the mean square is the square of the average of the factor saturations, so that the standard deviation of the two components used in the analysis of variance is equal to the average factor saturation. Thus, for the first factor the square of the average factor saturation is $(3·5101 \div 4)^2 = ·8776^2 = ·7702$ (the mean square as given in Table VI); for the second factor it is $(1·5816 \div 4)^2 = ·3954^2 = ·1564$. With the later factors, owing to the differential weighting, the relations become more complicated.

we endeavour to estimate a set of criterion-measurements from a given set of observed measurements, whose correlation with the former is known ; and the same formula can be employed, viz. (with the usual notation¹) $P = F'R^{-1}M$. Accordingly we can go on to calculate the requisite regressions by post-multiplying the set of correlations between the criterion and the tests (i.e., F' , the column of factor saturations rewritten as a row) by the inverse (R^{-1}) of the square symmetrical matrix of intercorrelations between test and test (i.e., by the inverse of Table VII) : in case any reader desires to verify the results, the inverse is printed in Table IX.

TABLE IX. INVERSE OF CORRELATION MATRIX

Test	(1)	(2)	(3)	(4)
(1)	109.90944	-67.44976	-15.77384	-45.13488
(2)	-67.44976	42.77878	9.90965	26.78966
(3)	-15.77384	9.90965	5.42803	3.74890
(4)	-45.13488	26.78966	3.74890	22.25723

The regressions for the several factors (as estimated by simple summation) are shown in Table X. With the first or general factor *the regressions for all four tests turn out to be exactly the same*. They are in fact equal to the reciprocal of the sum of the factor saturations ($1/3.5001 = 0.2849$). As in dealing with the analysis of variance therefore (cf. Table V, last line but one), the factor-measurements for the first factor, when obtained by simple summation with unity in the leading diagonal,² turn out to be a *straight sum or average* of the observed test-measurements : the division by 3.5001 is required merely to reconvert the sums to unitary standard measure.

TABLE X. MATRIX OF WEIGHTS FOR CALCULATING FACTOR-MEASUREMENTS BY SIMPLE SUMMATION

Factor				I	II	III	IV
(1) Arithmetic	..	..	..	.2849	.6266	1.6346	10.3337
(2) English	..	..	..	.2849	.6266	-1.3928	-6.3540
(3) Drawing	..	..	..	.2849	-.6382	-1.8708	-1.1200
(4) Handwork	..	..	..	.2849	-.6382	1.1560	-4.5206

If we base our calculations on the residual measurements for the several examinees, the same will hold good of the factor-measurements for the remaining factors. But this is no longer obvious from inspection of the regressions, since they are applied to the initial test-measurements, not to the residuals themselves.

¹ See below, footnote 1, p. 19. This, of course, is simply Gauss's determinantal procedure for combining a series of observed measurements in accordance with the principle of least squares ('Theoria combinationis observationum erroribus minimis obnoxia,' 1823, *Werke*, Bd. IV, pp. 1-53 ; cf. *Theoria motus*, 1909, art. 184). The method has been in regular use for quantitative determination in all branches of science ; but is most familiar to psychologists from Karl Pearson's early memoir (*Phil. Trans. Roy. Soc., A*, CLXXXVII, 1896, pp. 253f.) or from later papers largely based on it.

² In factorizing by simple summation it is more usual to insert in the leading diagonal, not the total test-variance (taken as 1.0), but the proportion of variance due to the postulated common factors only : and the common factors postulated generally comprise the smallest number that suffice to account for the observed intercorrelations. This maximizes the 'unique' factors, and somewhat complicates the correspondences between the factors obtained by summational analysis and by analysis of variance respectively. It does not, however, seriously affect the underlying analogies to which I have drawn attention. But in any particular case it will mean that a precise correspondence between the several 'factors' and the several 'sources of variation' assumed in analysis of variance cannot be taken for granted without a special consideration of the implications of the data.

(b) The Method of Weighted Summation

Weighting the Correlations before Summation. Factor saturations and factor-measurements secured by the foregoing simple procedure may suffice for most practical purposes. But in strict theory it is evident that, if we stopped here, we should be guilty of the same inconsistency as was pointed out above: for we have calculated both factor saturations and factor-measurements on the provisional assumption that all the tests are equally saturated; yet the results eventually reached show that in point of fact the saturations are different. For a more precise determination, therefore, as was argued above on page 15, we must have recourse to successive approximation.

But now it will be natural to suggest that the iterative process may also be applied to the correlation table instead of to the test-measurements, and labour consequently saved. And the mathematical reader will have no difficulty in satisfying himself¹ that this suggestion is theoretically valid. Thus the essential difference between the provisional method of calculating the saturations by simple summation, described above, and the more rigorous procedure needed to secure more accurate estimates will consist simply in the fact that, for the latter purpose, we shall *first weight each row of inter- and self-correlations* with the trial-value for the corresponding saturation coefficient, before we add the figures by columns—very much as we previously proposed to weight each row of test-marks. The improved saturations for the several factors, reached in this way, are shown in Table XI.

TABLE XI. FACTOR SATURATIONS OBTAINED BY WEIGHTED SUMMATION

Factor	I	II	III	IV	Sum of Squares
(1) Arithmetic	·9791	·1908	·0275	·0610	1·0000
(2) English	·7850	·6126	—·0869	—·0376	1·0000
(3) Drawing	·8310	—·4897	—·2640	—·0085	1·0000
(4) Handwork	·9148	—·2851	·2851	—·0252	1·0000
Sum of Squares	3·1021	·7328	·1593	·0058	4·0000
Contribution to Variance (as a percentage)	77·55	18·32	3·98	0·15	

The Composition of the Factors. I have already argued that, however obtained, the factors are to be regarded as *weighted sums* of the test-marks or of their residuals. In the case of the measurements for the first factor this is easily seen. But in the case of the supplementary or bipolar factors this interpretation is not so readily grasped. It may therefore be instructive to pause for a moment, and note how in theory (though not in practical calculation) the measurements for a bipolar factor are, in point of fact, built up.

In Table XII the top line (numbered 1 in the right-hand margin) gives the factor-

¹ The proof is as follows. Let M be the matrix of test-measurements in unitary standard measure. Then M can always be expressed in the form $LV^{\frac{1}{2}}P$, where L and P are orthogonal matrices and V a diagonal matrix of factor-variances. Let R denote the correlation matrix and $F=LV^{\frac{1}{2}}$ the matrix of factor saturations. Then $R=MM'=LVL'=FF'$. And $F'R=V^{\frac{1}{2}}L'LVL'=VF'$. Thus $F'=V^{-1}(F'R)$, and $V^2=V^{\frac{1}{2}}L'LVL'LV^{\frac{1}{2}}=F'RF$. And, if F_1 denotes the trial-values and we write $F_2=V^{-1}F_1R_1$, a simple proof of convergence will show that F_2 is an improved estimate of F . V and L in fact are the 'latent roots' and 'latent vectors' of the correlation matrix R ; and the procedure described in the text is merely one of the recognized methods for calculating such constants: (see, e.g., Fraser, R. A., Duncan, W. S., and Collar, A. R., *Elementary Matrices* (1938), pp. 140f.).

measurements for the general factor. Down the marginal column to the left the first four figures, in brackets under the name of each test, state its saturation for that factor. To obtain the contributions of the first factor to the several tests we have to multiply each boy's factor-measurement (as shown at the top) by these saturations. The products are inserted in lines 3, 6, 9, and 12. The test-marks themselves are placed in the rows above (lines 2, 5, 8, and 11); and from these initial marks we accordingly subtract the contributions of the first factor. Lines 4, 7, 10, and 13 show the remainders or 'residual marks.' The fact that we have in this way *eliminated from each test the whole of the first factor* can be verified by correlating each row of residuals with the measurements for the first factor at the top: in every case, it will be found, the *correlation is zero* (apart, of course, from trifling discrepancies due to rounding). Further, if the reader will correlate any one row of residuals with any other, he will find that the product-sum is identical with the residual correlation left after deducing the effects of the first factor from the observed correlation: (e.g., the product-sum of lines 4 and 8 is $\cdot 1122$: and the residual correlation for Arithmetic and English, calculated from Table VII, is $\cdot 8808 - \cdot 9791 \times \cdot 7850 = \cdot 1122$). Evidently then we can base our calculation of the second factor saturations either *on the residual marks themselves* or (what is plainly quicker) *on the residual correlations*.

TABLE XII. CALCULATION OF FACTOR-MEASUREMENTS FOR SECOND FACTOR

	John	James	Hugh	George	Ralph	Henry	Line
1st Factor-measurement	$\cdot 7049$	$\cdot 1068$	$\cdot 0000$	$-\cdot 0313$	$-\cdot 0850$	$-\cdot 6954$	(1)
(1) Arithmetic ..	$\cdot 6713$	$\cdot 2517$	$\cdot 0000$	$-\cdot 1678$	$-\cdot 0839$	$-\cdot 6713$	(2)
(.9791) ..	$\cdot 6893$	$\cdot 1043$	$\cdot 0000$	$-\cdot 0305$	$-\cdot 0832$	$-\cdot 6809$	(3)
Residual ..	$-\cdot 0190$	$\cdot 1474$	$\cdot 0000$	$-\cdot 1373$	$-\cdot 0007$	$\cdot 0096$	(4)
(2) English..	$\cdot 5941$	$\cdot 4951$	$\cdot 0000$	$-\cdot 3961$	$-\cdot 2970$	$-\cdot 3961$	(5)
(.7850) ..	$\cdot 5533$	$\cdot 0838$	$\cdot 0000$	$-\cdot 0246$	$-\cdot 0666$	$-\cdot 5459$	(6)
Residual ..	$\cdot 0408$	$\cdot 4113$	$\cdot 0000$	$-\cdot 3715$	$-\cdot 2304$	$\cdot 1498$	(7)
(3) Drawing ..	$\cdot 5307$	$-\cdot 1516$	$\cdot 0000$	$\cdot 4549$	$-\cdot 1516$	$-\cdot 6823$	(8)
(.8310) ..	$\cdot 5859$	$\cdot 0888$	$\cdot 0000$	$-\cdot 0260$	$-\cdot 0707$	$-\cdot 5779$	(9)
Residual ..	$\cdot 0552$	$-\cdot 2404$	$\cdot 0000$	$\cdot 4809$	$-\cdot 0809$	$-\cdot 1044$	(10)
(4) Handwork ..	$\cdot 6799$	$-\cdot 1943$	$\cdot 0000$	$\cdot 0000$	$\cdot 1943$	$-\cdot 6799$	(11)
(.9148) ..	$\cdot 6448$	$\cdot 0977$	$\cdot 0000$	$-\cdot 0286$	$-\cdot 0777$	$-\cdot 6362$	(12)
Residual ..	$\cdot 0351$	$-\cdot 2920$	$\cdot 0000$	$\cdot 0286$	$\cdot 2720$	$-\cdot 0437$	(13)
Residuals Weighted by Second Factor Saturations							
Test Weights							
(1) (.1908) ..	$-\cdot 0036$	$\cdot 0281$	$\cdot 0000$	$-\cdot 0262$	$-\cdot 0001$	$\cdot 0018$	(14)
(2) (.6126) ..	$\cdot 0250$	$\cdot 2520$	$\cdot 0000$	$-\cdot 2276$	$-\cdot 1411$	$\cdot 0918$	(15)
(3) (—·4897) ..	$\cdot 0270$	$\cdot 1177$	$\cdot 0000$	$-\cdot 2355$	$\cdot 0397$	$\cdot 0511$	(16)
(4) (—·2851) ..	$-\cdot 0100$	$\cdot 0832$	$\cdot 0000$	$-\cdot 0082$	$-\cdot 0776$	$\cdot 0125$	(17)
Weighted Sum ..	$\cdot 0384$	$\cdot 4810$	$\cdot 0000$	$-\cdot 4975$	$-\cdot 1791$	$\cdot 1572$	(18)
2nd Factor-measurement	$\cdot 0524$	$\cdot 6564$	$\cdot 0000$	$-\cdot 6789$	$-\cdot 2445$	$\cdot 2145$	(19)
Average Difference	$\cdot 0000$	$\cdot 6708$	$\cdot 0000$	$-\cdot 6708$	$-\cdot 2236$	$\cdot 2236$	(20)

The '2nd Factor measurement' is obtained from the 'weighted sum' by dividing the latter by $\cdot 7328$ (the sum of the squares of the saturations or 'weights').

The factor-measurements for the second factor consist of a *weighted sum of these residuals*. The weights used will be proportional to the saturations for the second factor (obtained by successive approximation). The weighted residuals themselves are set out near the foot of the table in lines 14 to 17, the saturations being given as before in brackets on the left. Line 18 gives the weighted sum. To bring these sums to unitary standard measure we must divide by the square-sum of the factor saturations ($\cdot 7328$). The standardized factor-measurements are given in the last line but one (line 19).

Curiously enough, the unstandardized figures (in the last line but two) can also be obtained¹ *direct from the observed test-measurements* by multiplying them by the same saturations: this again the reader should test for himself. In practice this more direct procedure furnishes the most convenient working-method for computing the factor-measurements in standard form, since the factor saturations can first be divided by their own variance before being used as weights. Finally, on correlating the second factor-measurements with the marks for each of the four tests, the reader can verify that the correlation of the factor with each test is identical with the saturation for that test as calculated by the usual short-cut from the intercorrelation table.

The measurements for the second factor should now be compared with the figures obtained by 'simple averaging' (designated the 'Average Difference' in the last line of Table II above). For purposes of comparison these are reduced to standard measure and appended at the foot of Table XII. The agreement is here remarkably close. Apart from the fact that in the fuller procedure just described the numerical weights are different instead of equal, the method used in Table II was essentially the same. In both cases the general factor is deducted from the observed marks: the first two rows of residuals are then weighted positively, the second two negatively, and all four then summed algebraically for each examinee.

The Use of Orthogonal Factors. Whether obtained by simple or by weighted summation, the special characteristic of the factor-measurements is that *the measurements for any one factor are uncorrelated with those for any other*. This (as we saw at the outset) does not hold good of the rough and ready factor-measurements obtained by averaging marks or differences between marks without any weighting at all: their intercorrelations will be low, especially with large samples of persons and tests, but not (except by accident) zero. One advantage of the factorial procedure, therefore, is that we are thus enabled to transform the n correlated sets of test-measurements into n uncorrelated sets of factor-measurements.

The idea of reducing a number of correlated variables to an equal number of uncorrelated variables, by choosing the 'principal axes of the correlation ellipsoid' to represent the new 'dimensions,' appears to have been first put forward by Karl Pearson.² As I have pointed out above, this procedure amounts to seeking in succession

¹The reason is that, with weighted summation as we have seen, the product-sum of the saturations for the second factor with those for the first factor is zero: hence multiplying by the former saturations automatically eliminates the first factor.

²*Biometrika*, I, 1901, p. 209. Garnett later gave a formal proof that " n correlated variables can always be expressed as linear functions of n uncorrelated variables" (*Brit. J. Psych.*, X, 1920, pp. 242f.). This, as he points out, can be done in many ways; but a *unique* solution is obtained, if we then select as the 'highest common factors' those which yield the largest contributory factor-variances, as Pearson's method requires. In previous papers I have shown that the procedure thus proposed is equivalent to the reduction of the correlation matrix to a classical 'canonical form.' For a more detailed discussion of this point and of the proof given in the previous footnote, and its various consequences, cf. Burt, C., *Brit. J. Psych.*, XXVIII (1937), pp. 77f., and *Psychometrika* III (1938), pp. 153f.

those hypothetical factor-measurements which give the closest fit to the observed test-measurements (and to the successive residuals) as judged by the principle of least squares. The reader will find it instructive to test the fit for himself by calculating the squared residuals left after deducting the factor-measurements for the first factor when calculated by the different methods enumerated in this article. He will find that the sum of the squares of the discrepancies is as follows : I (a) (Simple Averaging with Unstandardized Marks, divided by 10 to make the figures comparable) 1.180 ; I (b) (Simple Averaging with Standardized Marks) 0.920 ; II (a) (Factor Analysis by Simple Summation) 0.899 ; II (b) (Factor Analysis by Weighted Summation) 0.873. The practical worker will note that, so far as the first factor is concerned, the most marked improvement is that effected by the preliminary standardization.

The Analysis and Reconstruction of the Test-measurement. The formulæ commonly given for deducing a person's factor-measurements from his test performances nearly all assume that the factor-measurements are required to be in standard measure (cf. Table XII). This inevitably makes their variances equal. But for many purposes it is more instructive to express the factor-measurements in a way which indicates their relative importance. In that case *the factor-variances must be treated as properties of the factor-measurements* rather than of the factor saturations.

Now it is a peculiarity of weighted summation that, for every pair of factors, the product-sum of the saturations (as well as of the factor-measurements) is necessarily zero. Hence, if each column of factor saturations is normalized (by dividing it by the square root of its square-sum), *the factor matrix will be reduced to an orthogonal matrix of weighting coefficients.* These 'test-weights' are shown in lines 5 to 8 on the left of Table XIII. They are, in fact, the direction cosines forming the rotation matrix which changes the axes of reference to the principal axes of the correlation ellipsoid. On weighting the test-marks by these normalized saturations, we shall obtain the factor-measurements in the form shown in the middle of Table XIII (lines 5 to 8). The variances of the several sets of factor-measurements are now widely different ; they are, in fact, equal to the square-sums of the original factor saturations. The factor-measurements as thus computed may therefore be taken as indicating *the final outcome of the factorial analysis.*

And now from these independent or orthogonal sets of factor-measurements we can, if we wish, reconstruct the initial test-measurements exactly. The mode of reconstruction is shown in the last four lines of Table XIII. Since the inverse of an orthogonal matrix is merely the same matrix transposed, we can use *the same set of weights* to reconstruct the test-measurements from the factor measurements as we used to deduce the factor-measurements from the test-measurements. By their means we can in effect rotate the factor-axes back to the original position of the test-axes.

IV. FINAL COMPARISON

The Correspondence of the Factors. The reader will now be in a position to compare in detail the results of the different methods of analysis. The practical teacher will be chiefly concerned to know how far our first rough assessments for general ability, special aptitudes and the like have been modified by these more complicated techniques. Accordingly, let us begin by comparing the first and simplest with the last and most elaborate of our four main methods of calculation. We started, it will be remembered, by averaging the raw marks for each pupil to obtain an assessment for his general ability, and then went on to average the differences in his marks for the two types of subject, for the two days, and so on. In that way we reached the figures set out in the central section of Table III. To compare those

TABLE XIII. ANALYSIS AND RECONSTRUCTION OF TEST-MEASUREMENTS WITH DIFFERENTIAL WEIGHTS OBTAINED BY WEIGHTED SUMMATION

Test				John	James	Hugh	George	Ralph	Henry	Sum of Squares
				Observed Marks in Unitary Standard Measure						
(1) Arithmetic	..	..		·671	·252	·000	—·168	—·084	—·671	1·0000
(2) English	..	..		·594	·495	·000	—·396	—·297	—·396	1·0000
(3) Drawing	..	..		·531	—·152	·000	·455	—·152	—·682	1·0000
(4) Handwork	..	..		·680	—·194	·000	·000	·194	—·680	1·0000
Total	..	..								4·0000
Test Weights				Analysis into Factor-measurements						
(1)	(2)	(3)	(4)							
I. ·556	·446	·472	·519	1·242	·188	·000	—·055	—·150	—1·225	3·1021
II. ·223	·716	—·572	—·333	·045	·562	·000	—·581	—·209	—·184	·7328
III. ·069	—·218	—·662	·714	·051	—·129	·000	—·226	·298	·006	·1593
IV. ·798	—·492	·112	—·330	—·040	·038	·000	·010	·032	—·040	·0058
Total	..	..								4·0000
Factor Weights				Reconstruction of Test-measurements						
I	II	III	IV							
(1) ·556	·223	·069	·798	·671	·252	·000	—·168	—·084	—·671	1·0000
(2) ·446	·716	—·218	—·492	·594	·495	·000	—·396	—·297	—·396	1·0000
(3) ·472	—·572	—·662	·112	·531	—·152	·000	·455	—·152	—·682	1·0000
(4) ·519	—·333	·714	—·330	·680	—·194	·000	·000	·194	—·680	1·0000
Total										4·0000

TABLE XIV. ANALYSIS INTO FACTOR-MEASUREMENTS (BY ORDINARY AVERAGING, REDUCED TO COMPARABLE FORM)

Factor				John	James	Hugh	George	Ralph	Henry	Sum of Squares
I. General Ability				1·223	·175	·000	·000	—·175	—1·223	3·0535
II. Academic v. Technical Subjects				·000	·524	·000	—·524	—·175	·175	·6107
III. Quantitative v. Aesthetic Subjects				·087	—·087	·000	—·175	·262	—·087	·1221
IV. Monday v. Tuesday Subjects				·087	—·087	·000	·349	—·087	—·262	·2137

figures with our final results, it will be well to reduce them to terms of the same total square-sum or variance, viz., 4·0000. This we can do by simply multiplying them by $\sqrt{4 \div 131} = 0·17474$. The average marks, rescaled in this way, are shown in Table XIV. Factor I, it will be remembered, we regarded as denoting general ability in all four subjects; Factor II, superiority in the academic subjects as contrasted with the practical; Factor III (somewhat doubtfully), superiority in the æsthetic subjects as contrasted with the quantitative; and Factor IV, superiority in the subjects tested on Monday as contrasted with those tested on Tuesday. The figures in Table XIV are now to be compared with those in the central panel of Table XIII. For the first three factors *there is very little difference between the measurements obtained by simple averaging and those obtained by factor analysis*: with one trifling exception the order of the boys for each remains the same. Only with Factor IV (influence of days) is

there any considerable alteration ; and here, as we have seen, the revised factor-measurements are reduced virtually to vanishing quantities.

The factor-measurements obtained with the other two methods—averaging standardized marks and factor analysis by simple summation—will fall between the two extreme cases we have just compared ; and the divergences will consequently be smaller still.

A more detailed study would turn on a comparison, for each method in turn, of three distinctive features : (i) the normalized weighting matrices ; (ii) the normalized factor-measurements ; and (iii) the factor variances. Of these the first is crucial, since the rest are incidental consequences of the particular weights employed to effect the proposed transformation. The reader therefore should specially note the resemblances and differences between the sets of weights shown in Table XIII and those shown above in Table III for 'summation with *equal* weights' : those in Table III, it will be remembered, represent the system of weighting implicitly adopted in the method of simple averaging and in ordinary analysis of variance, except that, for the latter purpose, the 'test-weights' taken would be $1/\sqrt{n}$ instead of $1/n$, i.e., 0.50 instead of $\frac{1}{4}$. For Factor I, it will be seen, the *differential* test-weights shown in Table XIII do not differ widely from 0.50. But for the remaining factors the divergences become more marked, though (except for the last factor) the original pattern of signs is still preserved.

The Significance of the Factors. Now, if the parallel I have been drawing between factor analysis and the analysis of variance is justified, it would follow that we may adopt the method used in the latter type of analysis to test the significance of the factors, viz., comparing the mean-squares calculated from the several factor-variances, having due regard to the degrees of freedom involved. Hitherto tests of significance in factor analysis have generally been applied either to the residuals of the inter-correlations or (more rarely) to the factor saturations. But I suggest that it is rather *the factor-measurements themselves* that call primarily for such a test. The variances of these factor-measurements we have already calculated in the last column of Table XIII (lines 5 to 8). It remains to allocate the degrees of freedom.

Degrees of Freedom. We started with 4 rows of 6 measurements, in all 24 independent observations. But in reducing the measurements observed to standard form, we had to calculate both the means and the standard deviations from the data themselves.¹ Hence the standardized measurements (as set out in Table XV) involve only 4 degrees of freedom in each row, i.e., 16 in all. On calculating the factor-measurements, by taking weighted sums of the standardized test-marks, we reach a set of figures, each row of which must by hypothesis have zero correlation with the preceding rows. As a result each row after the first will lose 1, 2, and 3 degrees respectively. On the other hand, by choosing weighting coefficients that will maximize the variances of each row in order, we tend to magnify the differences in the earlier rows. The 4×4 rotation matrix, which contains the weighting coefficients, involves 6 degrees of freedom calculated from the data. Since each row of weights is orthogonal to the preceding rows, and every row must have a square-sum of one, the first row

¹ As noted above, Dr. M. S. Bartlett (to whom I am much indebted for comments on this point) would neglect the effect of fixing the standard deviation. His allocation of the degrees of freedom would therefore be $8+6+4+2=20$. As he points out, the problem is not unlike that arising in multivariate analysis (with an external criterion) where the partition of a sample can be regarded as a regression analysis, and treated by means of a canonical reduction. The effect exercised on the degrees of freedom by the choice of maximal weighting coefficients is itself part of a more general problem which seems to deserve further examination. For a discussion of the distributional theory of a factorial analysis into principal components, the reader may refer to Wilks, S. S., *Mathematical Statistics* (1943).

must involve 3 degrees, the next 2 degrees, and the last only 1 degree of freedom. These additional degrees must therefore be assigned to the weighted sums. Accordingly our final allocation of the degrees of freedom will be as follows: (i) first row of factor-measurements contains $6-2+3=7$ degrees; (ii) second row, $6-2-1+2=5$ degrees; (iii) third row, $6-2-2+1=3$ degrees; (iv) last row, $6-2-3+0=1$ degree: total, $7+5+3+1=16$.

Analysis of Factor Variance. The mean squares and variance-ratios¹ calculated on this basis are shown in Table XV. A comparison of this final table with Table VI above will show the analogies and the differences between the procedure here proposed and that of an ordinary analysis of variance. As before, we may begin by taking the fourth factor as providing our estimate of error for testing the third factor. And then, proceeding by progressive pooling as before,² we should eventually find that only the first factor is significant.

TABLE XV. ANALYSIS OF VARIANCE: (c) WITH FACTORS OBTAINED BY WEIGHTED SUMMATION

Factor	Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Critical Ratio	Probability
I	Between Persons	7	3.1021	.4432	4.44	0.01-0.05
II	Between Persons and Subjects	5	.7328	.1466	(3.55)	0.05-0.20
IV	Between Persons and Days	3	.1593	.0531	(9.16)	0.05-0.20
III	Residual	1	.0058	.0058	—	—
	Total	16	4.0000	—	—	—

¹ Since the partitioning of the total variance is now made after the observed data have been transformed by weighting equations deduced from the intercorrelations of the tests, the mode of analysis exemplified in Table XV can only be described as an 'analysis of variance' in a somewhat broadened sense of that phrase. How far (if at all) these preliminary transformations affect the strict applicability of the ordinary tables of variance-ratios is perhaps a question for further discussion. But the procedure proposed can at least be defended as a reasonable *approximate* test. Dr. Finney (in a personal letter commenting on a roneo'd report of mine where this procedure was employed) reminds me that frequent use is made of somewhat similar approximate tests in applying the chi-squared procedure.

In the present example, the correspondence between the 'factors' obtained by factor analysis and what would usually be described as the 'sources of variation' in analysis of variance (viz., variations 'between persons,' 'between persons and subjects,' etc.) seems to be fully justified; and a similar correspondence would, I believe, hold good in the majority of empirical problems for which factor analysis has hitherto been used. Yet cases are readily conceivable in which the correspondence might break down, and in which some of the factors might even represent components with no plausible meaning whatever. Hence in any concrete inquiry it is incumbent on the investigator to consider whether the underlying postulates are duly fulfilled, and, in particular, whether the weighting coefficients do sort out the sources of variation in the way assumed: but such precautions are, of course, essential in every factorial study.

² I must again remind the reader that, owing to the artificial nature of the small selected mark list chosen for examination, the figures in the table are illustrative only. Hence the 'probabilities' in the last column must be accepted merely as indicating the general scheme of tabulation. The detailed steps in the progressive pooling (a procedure, of course, which may not always be warranted) are as follows. The figure 5.33 (in brackets) is the ratio of the mean-square 'between persons and days' to the mean-square 'residual.' This proves to be non-significant. (Some caution is required at this stage: if, as here, the initial marks have been rounded off at the outset, a degree of freedom may have been lost in the process; on the other hand, if the saturations have been obtained by successive approximation, the final 'residual' may be swollen by the imperfections of that approximation, and what should strictly be zero may perhaps appear as a misleading small residual.) Since their ratio has proved to be non-significant, the two smallest mean-squares are next pooled to test the mean-square 'between persons and subjects.' The resulting ratio (3.52) also turns out to be non-significant. Accordingly, all three are now pooled to test the mean-square 'between persons.' And this final test here happens to lead to an apparently significant result. On the caution to be observed in adopting this pooling procedure, see above, p. 11.

V. SUMMARY AND CONCLUSIONS

1. A simplified table of examination marks, obtained with four tests, has been taken as an illustrative example ; and the data analysed by four main methods : (i) by the common-sense method of simple averaging, (ii) by an analysis of variance, (iii) by factor analysis with simple summation, and finally (iv) by factor analysis with weighted summation. Certain instructive relations and analogies have been elicited, which indicate, first, that an ordinary analysis of variance may be used to obtain much the same factors as are obtained by factor analysis, and secondly, that the results of factor analysis may be tested by the procedure adopted in an analysis of variance.

2. It is first shown that (a) the common-sense method of averaging the observed marks and then averaging the successive deviations and (b) the extraction of the components whose variances are calculated for an analysis of variance are both equivalent to pre-multiplying the matrix of test-marks by a quasi-orthogonal weighting matrix containing equal weights, but different signs ; and that the marks can be reconstructed from these components by the inverse of this weighting matrix. The procedure is thus formally similar to the use in factor analysis of a matrix of regression coefficients to obtain factor-measurements from test-measurements, and a matrix of factor saturations to construct test-measurements from factor-measurements.

3. It is then shown that with analysis of variance the weighting-matrices are orthogonal, but the factor-measurements are not. With factor analysis by simple summation, the weighting matrices are not orthogonal, but the factor-measurements are. With factor analysis by weighted summation *both* the weighting matrices and the factor-measurements are orthogonal ; and, as in the analysis of variance, the weighting matrix used to reconstruct the test-measurements from the factors is simply the transpose of the weighting matrix used to obtain the factor-measurements from the tests.

4. The analysis of variance presupposes a more explicit prior formulation of the hypothetical partitioning of the data into factors, and yields more definite tests of significance. With analysis of variance, however, as with simple averaging, all the weights are numerically equal : factor analysis, on the other hand, while preserving much the same pattern of signs, weights the items differentially. As a consequence, factor analysis yields weighting matrices which have the double merit of furnishing factor-measurements which are not only (i) uncorrelated, but also (ii) take into account the differing value of the several test-measurements as aids in estimating the hypothetical factor-measurements.

5. Nevertheless, the factor-measurements obtained by factor analysis are, in their fundamental nature, still simply weighted sums or averages. And it is shown by actual examples that the same factor-measurements can be obtained by the method of weighted averaging, applied direct to the observed marks, without calculating correlations. The chief reason for starting with a preliminary calculation of the correlations is that this device shortens the working-procedure. The bipolar factors so obtained are then found to correspond with what in analysis of variance are termed interactions : this confirms the view that bipolar factors may have a meaning as they stand, and involve a reciprocity between tests or traits and persons.

6. Finally, it has been shown that the significance of the factors can be tested, at least approximately, by the method adopted in an ordinary analysis of variance, provided the degrees of freedom are re-allocated to allow for the effects of standardization and of differential weighting.

THE MAXIMUM CORRELATION OF TWO WEIGHTED BATTERIES (HOTELLING'S "MOST PREDICTABLE CRITERION")

By GODFREY THOMSON

Moray House, Edinburgh

I. *Introduction.* II. *Mathematical Treatment.* III. *A Numerical Example as Model.*
IV. *Exchanging Battery and Criterion.*

I. INTRODUCTION

In accordance with the suggestion that some articles in this Journal should aim at interpreting mathematical methods to readers who are not professional mathematicians but have had some statistical training, the present article takes Hotelling's idea (1 and 2) of weighting separately two different batteries of tests or other variables in such a way that the two weighted sums correlate as highly as possible with one another; translates it into a matrix notation perhaps more familiar to our readers; and works out in complete detail a model example, to which those who are quite non-mathematical should turn after reading the next three paragraphs. The model, carefully followed, should enable them to apply the method.

Ordinarily we predict, i.e., estimate, a *single* criterion (such as marks in a future examination, or degree of success in a vocation) by means of a battery of several tests, including under "tests" not only group or individual tests, but ratings at interviews, or any scores which influence the prediction. The best weights to use on the individual tests of the battery are then the regression coefficients calculated in the usual way.

The criterion might conceivably however be composite, indeed it usually is really composite, being formed of marks in different subjects in a School Leaving Certificate, or of opinions by different foremen on the same apprentice. These however are generally condensed into one mark called *the* criterion by giving the component parts different weights, or possibly equal weights. These weights, though they are usually I fear almost accidental and sometimes unconscious, can be conceived as chosen to give that combination of the components which in the opinion of good judges best represents success at the school, or in the occupation. In that case the criterion is again reduced to a single measure, and the regression coefficients of the tests of the battery are calculated in the usual way to give multiple, that is maximum, correlation or prediction.

Hotelling's idea, however, is to give weights to the parts of the criterion which, like the weights given to the parts of the battery, will cause the combined sum of the one to correlate as highly as possible with the combined sum of the other. At this point the distinction between criterion and battery breaks down. They are simply two batteries which have to be weighted—the weights will be in general different for the two batteries—so that they agree as closely as possible with one another. Situations in psychological work where this is desirable will readily occur to the reader. I have already (3) applied the idea to a very special case, where the two batteries are composed of Forms A and B of the same tests and the Hotelling procedure leads to the best weights to maximize battery reliability. Here, however, we consider the general case of two different batteries, composed of different tests

and not necessarily the same number of tests. In our later example I have taken three tests in the one battery and two in the other.

Before we go on to any algebraic treatment it may, for some readers, be illuminating to consider the problem as a geometrical one. If there are, say, three tests in the one and two in the other battery, the vectors or lines representing these five tests exist in a "space" of five dimensions, each test-vector being at angles with the others which represent by their cosines its correlations with them. The one battery is in a three-dimensional space, the other in a two-dimensional space or plane. This plane is not, in general, within that three-dimensional space: if it were, the two batteries could by proper weighting be made to agree exactly. The problem is so to weight the two tests in the plane that their resultant is as near as possible to the three-dimensional space, and similarly so to weight the three tests of the other battery that their resultant is as near as possible to the plane. More exactly, we want the weighted resultant of the three to make as small an angle as possible with the weighted resultant of the two.

II. MATHEMATICAL TREATMENT

Let the internal correlations, and the cross correlations of the tests of the two batteries be symbolized as follows:

	weights a'	b'
weights a	A	C'
weights b	C	B

Here A stands for the whole square matrix of correlations of the first battery, and B for those of the second battery. Thus if there are three tests in the first battery,

$$A = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

C similarly represents the cross-correlations of the tests of the first battery with those of the second. For example, with two tests in the second battery, we would have

$$C = \begin{bmatrix} r_{14} & r_{24} & r_{34} \\ r_{15} & r_{25} & r_{35} \end{bmatrix}$$

The matrix C' is the same turned round so that rows and columns become columns and rows. This is called *transposing* a matrix.

The symbol a represents all the weights given to the tests of the first battery, i.e., if there are three tests, a stands for the column of weights

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

while a' stands for the row a_1, a_2, a_3 .

Similarly with the symbol b . Our object is to find weights a and b which will maximize the correlation between the two batteries.

When the B battery consists of only one test the weights to be applied to the A battery are

$$a = A^{-1}r \quad (1)$$

where r is the single column of correlations to which C' has been reduced. The multiple, i.e., maximum possible correlation R is given by

$$R^2 = r' A^{-1}r \quad (1a)$$

This is the usual case of a single "criterion."

If there are several tests in the B team, with weights b , the correlations of the B team, thus weighted, with the separate tests of the A team will be given by

$$r = \frac{C'b}{\sqrt{b'Bb}} \quad (2)$$

We shall call the operation represented by eqn. (2) "condensing the B team." It reduces the matrix

$$\begin{bmatrix} A & C' \\ C & B \end{bmatrix} \quad \text{to} \quad \begin{bmatrix} A & r \\ r' & 1 \end{bmatrix}$$

If we substitute from (2) in (1a) we have

$$R^2 = \frac{b'CA^{-1}C'b}{b'Bb} = \lambda \quad \text{say.} \quad (3)$$

This has to be made a maximum by the proper choice of weights b . Therefore we must have

$$dR^2/db = 0 \quad (4)$$

$$\text{i.e., } b'Bb \frac{d}{db} (b'CA^{-1}C'b) = b'CA^{-1}C'b \frac{d}{db} (b'Bb)$$

$$\text{i.e., } b'Bb \times CA^{-1}C'b = b'CA^{-1}C'b \times Bb$$

$$\text{i.e., } (CA^{-1}C' - \lambda B) b = 0 \quad (4a)$$

a set of homogeneous equations in b . We must therefore have

$$|CA^{-1}C' - \lambda B| = 0 \quad (5)$$

an equation for λ . And since $\lambda=R^2$ which we want to maximize, it is the largest root we want. This root can be found by trial and error, calculating the above determinant for λ =the square of the largest correlation among the cross correlations in C , and working upward till the sign changes, then interpolating. But it would be very much better if some quicker and less laborious way of finding λ could be devised.

Equation (4a) then gives the weights b in ratio, namely, they are proportional to the co-factors of any row of $(CA^{-1}C' - \lambda B)$. The weights a are then found by condensing the B team, standardizing it, and performing the usual regression calculation. The whole calculation can be repeated reversing the rôles of the A and B teams, with the same result.

III. A NUMERICAL EXAMPLE AS MODEL

Let the correlations be

1.0	0.1	0.6	0.7	0.2	=	<table><tr><td>A</td><td>C'</td></tr><tr><td>C</td><td>B</td></tr></table>	A	C'	C	B
A	C'									
C	B									
0.1	1.0	0.4	0.3	0.8						
0.6	0.4	1.0	0.5	0.3						
0.7	0.3	0.5	1.0	0.4						
0.2	0.8	0.3	0.4	1.0						

The first step is to find $C A^{-1} C'$ by Aitken's pivotal condensation. It is

$$\begin{bmatrix} 0.5434 & 0.3210 \\ 0.3210 & 0.6693 \end{bmatrix}$$

Equation (5) is then

$$\begin{vmatrix} 0.5434-\lambda & 0.3210-0.4\lambda \\ 0.3210-0.4\lambda & 0.6693-\lambda \end{vmatrix} = 0$$

As this is only a quadratic it can be easily solved. To keep the example general however proceed by trial and error. The largest correlation in C is 0.8 so λ must be greater than 0.64.

Try $\lambda = 0.67$ deter. = -0.00281
 $\lambda = 0.70$ deter. = 0.00303 (plus)

Linear interpolation gives $\lambda = 0.6845$, $R = 0.83$
 $(CA^{-1}C' - \lambda B)$ b is then

$$\begin{bmatrix} -0.141 & 0.047 \\ 0.047 & -0.015 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = 0$$

whence $b_1 : b_2 :: 1 : 3$ very nearly,

(i.e., 15 to 47, or 47 to 141).

Now condense the B team thus, multiplying the B columns and rows by these weights :—

				1	3
	1.0	0.1	0.6	0.7	0.2
	0.1	1.0	0.4	0.3	0.8
	0.6	0.4	1.0	0.5	0.3
1	0.7	0.3	0.5	1.0	0.4
3	0.2	0.8	0.3	0.4	1.0

becomes

1.0	0.1	0.6	1.3	(= $0.7 + 3 \times 0.2$)
0.1	1.0	0.4	2.7	
0.6	0.4	1.0	1.4	
1.3	2.7	1.4	12.4	

Divide the last row and column by $\sqrt{12.4}$ and we get

1.0	0.1	0.6	0.105
0.1	1.0	0.4	0.218
0.6	0.4	1.0	0.113
0.105	0.218	0.113	1.000

from which can be found the A team weights by a regression calculation by pivotal condensation (4, pages 89 and 104) giving

or 0.107 0.224 -0.041
 2.61 5.47 -1.00

Maximum Correlation of Two Batteries

Thus the whole set of weights is

$$\begin{array}{ccccc} a_1 & a_2 & a_3 & b_1 & b_2 \\ 2.6 & 5.5 & -1.0 & 1.0 & 3.0 \end{array}$$

This can be checked by a "pooling-square" (4, chapter vi) thus :

	2.61	5.47	-1.00	1.00	3.00
2.61	1.00	0.10	0.60	0.70	0.20
5.47	0.10	1.00	0.40	0.30	0.80
-1.00	0.60	0.40	1.00	0.50	0.30
1.00	0.70	0.30	0.50	1.00	0.40
3.00	0.20	0.80	0.30	0.40	1.00

Multiply rows and columns and add the quadrants together, giving

$$\begin{array}{r|l} 33.0804 & 16.7620 \\ \hline 16.7620 & 12.4000 \end{array}$$

$$R = \frac{16.762}{\sqrt{33.08 \times 12.4}} = 0.83 \text{ as expected.}$$

IV. EXCHANGING "BATTERY" AND "CRITERION"

It is instructive now to repeat the whole calculation reversing the rôles of the teams. Writing the original correlations with the B team first

$$\begin{array}{c|c} B & C \\ \hline C' & A \end{array}$$

equation (5) is now

$$| C'B^{-1}C - \lambda A | = 0$$

Find first $C'B^{-1}C$

It is

$$\begin{bmatrix} .498 & .145 & .340 \\ .145 & .641 & .231 \\ .340 & .231 & .262 \end{bmatrix}$$

The equation for λ is therefore

$$\begin{vmatrix} .498 - \lambda & .145 - .1\lambda & .340 - .6\lambda \\ .145 - .1\lambda & .641 - \lambda & .231 - .4\lambda \\ .340 - .6\lambda & .231 - .4\lambda & .262 - \lambda \end{vmatrix} = 0$$

$$\lambda = .64 \text{ gives } +.00276$$

$$\lambda = .70 \text{ gives } -.00130$$

Linear interpolation then gives $\lambda = \cdot681$, $R = \cdot83$
($\cdot6845$ before)

The weights a are then proportional to any row of the adjugate of $(C'B^{-1}C - \lambda A)$
i.e., the adjugate* of

$$\begin{bmatrix} -\cdot187 & \cdot077 & -\cdot071 \\ \cdot077 & -\cdot044 & -\cdot043 \\ -\cdot071 & -\cdot043 & -\cdot423 \end{bmatrix}$$

which is

$$\begin{bmatrix} \cdot0168 & \cdot0356 & -\cdot00644 \\ \cdot0356 & \cdot0741 & -\cdot0135 \\ -\cdot00644 & -\cdot0135 & \cdot00230 \end{bmatrix}$$

any row being in the ratios approximately

$$\begin{array}{ccc} 2\cdot6 & 5\cdot5 & -1 \\ a_1 & a_2 & a_3 \end{array}$$

Now condense the A team as follows :

	2·6	5·5	-1·0		
2·6	1·0	0·1	0·6	0·7	0·2
5·5	0·1	1·0	0·4	0·3	0·8
-1·0	0·6	0·4	1·0	0·5	0·3
	0·7	0·3	0·5	1·0	0·4
	0·2	0·8	0·3	0·4	1·0

becomes

33·35	2·97	4·62
2·97	1·00	0·40
4·62	0·40	1·00

which becomes

1·00	0·089	0·139
0·089	1·000	0·400
0·139	0·400	1·000

* The adjugate of a matrix is composed of the co-factors of its terms. The co-factor of a term is the value of the determinant left when that term's row and column are deleted, with the appropriate sign, alternately plus and minus.

Maximum Correlation of Two Batteries

The b weights are then proportional to

	$\begin{vmatrix} \cdot089 & \cdot4 \\ \cdot139 & 1\cdot0 \end{vmatrix}$	and	$\begin{vmatrix} \cdot089 & 1\cdot0 \\ \cdot139 & \cdot4 \end{vmatrix}$
that is to			
	$\cdot0334$	and	$\cdot1034$
that is approximately			
	1	and	3
as by the former method.			

REFERENCES

1. Hotelling, H., Relations between two sets of variates, *Biometrika* 1936, XXVIII, 321.
2. Hotelling, H., The most predictable criterion, *J. Educ. Psychol.* 1935, XXVI, 139.
3. Thomson, Godfrey H., Weighting for battery reliability and prediction, *Brit. J. Psychol.* 1940, XXX, 357.
4. Thomson, Godfrey H., *The Factorial Analysis of Human Ability* (Univ. of London Press 1939 and 1946).

BIRTH ORDER, MATERNAL AGE AND INTELLIGENCE

WITH PARTICULAR REFERENCE TO A POSSIBLE EFFECT ON THE ASSOCIATION BETWEEN INTELLIGENCE AND FAMILY SIZE

By J. A. FRASER ROBERTS

Burden Mental Research Department, Stoke Park Colony, Bristol, and Department of Medical Statistics, London School of Hygiene and Tropical Medicine.

I. Introduction. II. Material. III. Comparison of Older and Younger Children born to the same Parents. IV. Family Size, Order of Birth and Intelligence in the Whole Group. V. Discussion. VI. Summary. Appendix.

I. INTRODUCTION

Advancing maternal age has been shown to be a factor of importance in determining the occurrence of certain abnormalities. The most striking example is mongolism, the incidence of which rather more than doubles with every five years advance in maternal age from 25 onwards (Penrose, 1938). A similar though less pronounced association has been found in a number of other conditions, for example, anencephaly, hydrocephaly, and congenital heart disease (Malpas, 1937). In at least one condition, congenital pyloric stenosis, it has been shown that order of birth is of much importance. Amongst those presumed to be of the appropriate genetic constitution, the first-born, if a boy, is about five times more likely to be affected than any girl, or than any subsequent child of either sex (Cockayne and Penrose, 1943). As would be expected, the possibility that advancing maternal age or advancing parity might have an effect on intelligence has often been investigated. There is also a considerable literature on the suggested slight handicapping of the first-born. A thorough and comprehensive review has been published by Jones (1933).

Advancing maternal age and advancing birth rank (i.e., place in the sibship) are necessarily highly correlated and when an association is demonstrated it is by no means easy to discover which factor is actually responsible. Complicated statistical methods will usually be necessary, for example the technique used by Penrose (1934), which enabled him to conclude that in mongolism it is advancing maternal age which is the responsible factor and that advancing birth rank is without effect. In intelligence, however, the problem is to determine whether there is any effect at all. Jones in the paper already quoted concludes that the association is not proved, and doubtless many other psychologists would agree with him. To test, therefore, whether a body of data reveals any significant association, it is sufficient to measure either maternal age, or, what is much more convenient in practice, birth rank. If either factor shows no significant association with intelligence, it can be concluded that the other factor also is without effect.

The problem, as regards intelligence, is of much interest and importance for its own sake. But there is another reason which makes it especially desirable to investigate any possible association. Many studies have shown a negative correlation between the intelligence of the child and the number of its sibs. These have recently

been reviewed by Burt (1947). It is therefore concluded by many that the average intelligence of the population must be declining. But if advancing maternal age or advancing birth rank are unfavourable factors it is clear that the lower average intelligence of the child from the large sibship might be due, at least in part, to the fact that it is a large sibship, and to that extent not to the supposed greater fertility of the less intelligent. If only the effect of maternal age or birth rank were great enough, though clearly it is not, the observed association between intelligence and the number of sibs could be accounted for completely in this way alone. There might theoretically be a negative correlation of the order observed between intelligence and sib number, but without any tendency for the less intelligent to be more fertile, and without any reason to suppose that average intelligence is declining.

The problem can be approached in two ways, both simple in theory, but involving considerable practical difficulties. The first method is to measure the intelligence, sib number, and birth rank of a large sample of schoolchildren and then by comparisons of the intelligence of children of different birth orders within each total size of family separately, determine whether there is any association. One difficulty is that sampling must be of children of school age, and therefore a proportion of sibships will be incomplete, the incompleteness itself being associated with the position of the child. Thus, last children in families of seven will tend to come from sibships which are more complete than will first children in families of seven. It is usually impossible to ascertain the age of the mother, as this would necessarily take the investigator from the school to the home, so no correction can be applied and the effect of total size of sibship cannot be satisfactorily eliminated. And there is always, of course, the difficulty of securing a sample which is not biased in relation to the variables concerned.

An alternative approach is to measure the intelligence of brothers and sisters born to the same parents in order to see whether there is any difference between the older and younger born. Here the main difficulty is quite different. The ideal would be to measure the children of each given sibship at the same chronological age, thus avoiding the necessity for age correction. In practice this is too difficult, and so the older born will also be older at test than the younger born. Hence correction of scores for difference in age at test must be quite extraordinarily precise. If the norms are only slightly and non-significantly too easy for the older children and slightly too difficult for the younger, there will be an apparent fall in intelligence from older to younger, or if the reverse is true, an apparent rise. In fact, it is difficult to avoid the conclusion that the only valid experiment must employ norms established on the group itself, so that relatively large numbers are required if anything less than a gross difference is to be detected. It would seem to be a permissible relaxation to establish norms on a large group of children, appropriately sampled, and then to apply those norms to such pairs of sibs as happen to be included. Studies of pairs of sibs in which published norms based on other material are applied cannot provide acceptable evidence. It is not sufficient to show that after applying such age corrections there is no significant residual association with age at test. It is shown later in this paper how a non-significant difference in the regressions from which norms were calculated turned a difference between the older and younger members of the pairs from something close to zero into an apparently highly significant difference in favour of the earlier born. In other words, in order to make a valid test of significance it is necessary that the regressions on which the norms are based must be unbiased estimates, and this is only possible if the regressions are established on the group itself, or, what comes to the same thing, if published norms are corrected by the results obtained on the group investigated. A further consideration which lessens the usefulness of this method of approach is that the difference in age between members of the pairs will usually be relatively small.

When all these difficulties are given proper weight, the number of researches which provide reasonably good evidence is not large. Accordingly it seemed desirable to add a further study by giving results obtained on data secured during a survey carried out by my colleagues and myself in the City of Bath, data which have already been used for a number of purposes (1935, 1937, 1938a, 1938b, 1940, 1945). Results are given, first, for a comparison of pairs of full sibs ; secondly, evidence is presented as to whether the observed association between intelligence and sib number depends to any significant degree on the effect of maternal age or birth rank.

II. MATERIAL

In selecting the sample of schoolchildren, an attempt was made to secure a complete cross-section of the population. The group was defined as all children whose homes were within the boundaries of the City of Bath on July 27th, 1934, and whose dates of birth fell between 1st September, 1921, and 31st August, 1925, inclusive. The total number was about 3,400. Owing to the excellent co-operation received ascertainment and mental testing were made almost complete. Details are given in the papers quoted, in particular, Roberts, Norman, and Griffiths (1938a, 1938b). The sample is clearly a suitable one upon which to establish norms for the test used, the Advanced Otis, though larger numbers would have been desirable in order to reduce the standard errors of the regressions on age at test to really small limits.

The sample included a number of sets of sibs born to the same parents, giving the equivalent of a comparison of 622 pairs of older and younger sibs. The first part of this paper deals with this comparison. Secondly, size of sibship was recorded for the whole group and also position in the sibship. The difficulty of incompleteness of families mentioned above remains, but it does seem possible to discover whether birth rank, and therefore alternatively advancing maternal age, did in fact contribute significantly to the association between intelligence and sib number which was described in the third paper of the series in the *Annals of Eugenics* (Roberts, Norman, and Griffiths, 1938a).

III. COMPARISON OF OLDER AND YOUNGER CHILDREN BORN TO THE SAME PARENTS

The sample included 504 pairs of sibs. There were also included 49 sets of three, each of which yields three comparisons, and two sets of four, each yielding six comparisons. The total number of 663 is, for the purpose of the present paper, reduced to 622. First, pairs of twins do not yield a comparison ; secondly, two or three children received an individual test only. Finally, five idiots and imbeciles were omitted ; this seems justifiable in view of the strong probability that there is a qualitative distinction which separates low-grade mental defectives from the rest of the population, a conclusion strongly reinforced by the findings in this sample, as was described in the fourth paper of the series (1938b). It seems desirable to exclude the low-grade defectives because the very large difference between them and their normal sibs would unduly increase the standard error of the difference between older and younger children.

Norms to allow for differences in age at test were originally calculated as was described in the first paper (1935). My colleagues and I have long been aware, however, that the estimates of the regressions on age were not strictly unbiased. This has mattered little up to the present because in no instance has any comparison involved age at test. In the present enquiry, however, the older born of each pair was on the average about eighteen months older at test than the younger. Accordingly, it was

necessary to recalculate the norms. This is described in the appendix to this paper. The recalculation also made it possible to base the norms on the whole sample instead of on three only of the four age groups, as was done originally.

Using the new norms of Table XI, each child contributing to the 622 comparisons was allotted a score which was simply the difference, plus or minus, between the observed and expected number. Table I shows the frequency distribution of the differences in these age-corrected scores between older and younger born children. The difference is called positive when the older sib is the cleverer and negative when the younger born is the cleverer. In order to compare first-born children with subsequent children the total frequency distribution is broken up to show, first, pairs which include a first-born child and, secondly, pairs not including a first-born child. It has been shown in this material that while the mean performance of boys and girls does not differ significantly, there is a greater variability amongst boys which is highly significant (Roberts, 1945). Accordingly, in Table I, boy pairs, girl pairs, mixed pairs with the boy older and mixed pairs with the girl older, are shown separately.

TABLE I. FREQUENCY DISTRIBUTION OF DIFFERENCE IN AGE-CORRECTED SCORE BETWEEN OLDER AND YOUNGER FULL SIBS

(+, older child cleverer)

	Pairs including first-born child	Pairs not including first-born child	Girl pairs	Boy pairs	Mixed pairs		Total
					Girl older	Boy older	
Group central difference in age-adjusted score*							
-105.5	1	—	—	1	—	—	1
- 95.5	—	—	—	—	—	—	—
- 85.5	1	1	—	1	1	—	2
- 75.5	5	3	1	2	3	2	8
- 65.5	3	5	1	4	3	—	8
- 55.5	9	7	8	4	1	3	16
- 45.5	11	20	8	9	4	10	31
- 35.5	10	35	7	11	15	12	45
- 25.5	22	35	16	13	15	13	57
- 15.5	17	36	14	16	11	12	53
- 5.5	28	43	20	23	16	12	71
+ 4.5	24	46	16	22	15	17	70
+ 14.5	32	35	20	21	12	14	67
+ 24.5	28	36	16	21	15	12	64
+ 34.5	20	32	16	11	12	13	52
+ 44.5	22	14	10	12	5	9	36
+ 54.5	15	13	5	8	6	9	28
+ 64.5	2	7	—	5	1	3	9
+ 74.5	1	1	—	—	1	1	2
+ 84.5	—	2	—	2	—	—	2
Total	251	371	158	186	136	142	622
Mean †	3.385	-0.379	0.576	2.296	-1.529	2.810	1.140
Standard error	2.146	1.639	2.394	2.511	2.797	2.788	1.307

* Top group, -110 to -101 ; 1st positive group, 0 to +9.

† Many of the calculations of this paper have been repeated using smaller units of grouping, but in no instance was the result appreciably different.

The mean difference for the total is 1.140 points in favour of the older born child, this being smaller than its standard error. The standard deviation of differences between age-corrected scores is 32.602, 28.6 times the standard error just given. Any temptation to argue that the difference, though non-significant, is at least in the expected direction is at once corrected by the division of the material into pairs including and not including a first-born child. All advantage in favour of the older members of the pairs is due to a superiority, which is not significant, of first born over subsequent children, and this is just what would not be expected. The comparison by sexes reveals nothing more than the known greater variability of boys. The apparent superiority of the older child when it is a boy is not significant. There is no indication, therefore, of any association between birth order and intelligence, though, of course, the numbers are not sufficient to exclude a relatively small effect.

The importance of unbiased regressions for the calculation of the norms is illustrated by the fact that although the new regression coefficients do not differ significantly from the old, as is shown in the appendix, nevertheless when the old norms were used the advantage in favour of the older born child exceeded three and a half points and was highly significant.

The pairs may be examined to see whether the difference increases with the interval between the dates of birth; Table II presents a correlation table. It also seemed desirable to examine the difference in relation to the age at test, as shown in Table III. The standard errors of the regressions of score on age are fairly large and it might have been found that adjustment for age at test within the 622 pairs would increase the superiority of the older born children. Difference in date of birth and difference in age at test are, of course fairly highly correlated (actually, $r=0.643$); so to complete the analysis Table IV shows the relation of these two variables.

TABLE II. SIB PAIRS. CORRELATION TABLE SHOWING DIFFERENCE IN DATE OF BIRTH AND DIFFERENCE IN AGE-CORRECTED SCORE

		Group central difference in date of birth (months)								Total
		11	16	21	26	31	36	41	46	
Group central difference in age-adjusted score	-105.5	—	1	—	—	—	—	—	—	1
	-95.5	—	—	—	—	—	—	—	—	—
	-85.5	—	1	1	—	—	—	—	—	2
	-75.5	—	3	2	1	—	2	—	—	8
	-65.5	1	3	1	2	—	1	—	—	8
	-55.5	1	2	5	3	1	3	—	1	16
	-45.5	3	9	7	5	1	2	4	—	31
	-35.5	4	11	9	12	2	5	1	1	45
	-25.5	3	11	12	12	8	8	2	1	57
	-15.5	1	12	16	10	3	6	4	1	53
	-5.5	8	14	21	10	11	3	4	—	71
	+ 4.5	1	13	21	16	9	7	1	2	70
	+ 14.5	4	15	12	9	16	5	6	—	67
	+ 24.5	3	15	18	14	9	3	2	—	64
	+ 34.5	1	13	12	11	6	7	2	—	52
	+ 44.5	2	5	11	5	6	2	4	1	36
	+ 54.5	—	8	8	5	3	2	2	—	28
	+ 64.5	—	4	2	3	—	—	—	—	9
	+ 74.5	—	1	1	—	—	—	—	—	2
	+ 84.5	1	—	—	1	—	—	—	—	2
Total		33	141	159	119	75	56	32	7	622

TABLE III. SIB PAIRS. CORRELATION TABLE SHOWING DIFFERENCE IN AGE AT TEST AND DIFFERENCE IN AGE-CORRECTED SCORE

	Group central difference in age at test (months)																		Total
	-5	-2	1	4	7	10	13	16	19	22	25	28	31	34	37	40	43	46	
-105.5	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	1
-95.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-85.5	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	-	-	2
-75.5	-	-	-	1	-	-	1	2	1	1	-	-	-	-	1	-	-	-	8
-65.5	-	1	1	-	-	-	-	2	-	-	1	3	-	-	-	-	-	-	8
-55.5	-	-	-	1	2	4	2	2	-	2	1	1	-	-	1	-	-	-	16
-45.5	-	1	1	3	-	-	3	4	4	1	4	4	3	2	1	-	-	-	31
-35.5	1	1	2	-	3	4	6	8	4	7	4	-	2	2	1	-	-	-	45
-25.5	-	1	-	3	-	4	2	7	13	8	8	4	2	1	2	1	1	-	57
-15.5	-	-	1	4	4	4	1	7	4	6	7	8	5	1	1	-	-	-	53
-5.5	1	-	1	6	8	10	3	4	12	6	8	5	5	2	-	-	-	-	71
+4.5	-	-	2	5	5	3	4	9	12	3	8	9	2	5	2	-	-	1	70
+14.5	-	1	2	1	8	6	5	7	7	10	8	3	5	1	2	1	-	-	67
+24.5	-	-	3	3	5	3	10	6	6	8	5	6	4	4	1	-	-	-	64
+34.5	-	-	1	4	5	4	2	5	5	5	7	9	2	3	-	-	-	-	52
+44.5	-	1	-	2	4	2	1	3	4	6	2	4	3	2	2	-	-	-	36
+54.5	-	-	1	-	3	2	-	2	5	5	2	3	3	2	-	-	-	-	28
+64.5	-	-	-	1	-	1	2	1	1	1	1	1	-	-	-	-	-	-	9
+74.5	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	-	2
+84.5	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	2
Total	2	6	15	34	47	47	44	72	78	70	68	60	36	25	14	2	1	1	622

TABLE IV. SIB PAIRS. CORRELATION TABLE SHOWING DIFFERENCE IN DATE OF BIRTH AND DIFFERENCE IN AGE AT TEST

	Group central difference in date of birth (months)								Total
	11	16	21	26	31	36	41	46	
Group central difference in age at test (months)									
-5	2	-	-	-	-	-	-	-	2
-2	3	2	-	-	1	-	-	-	6
1	1	13	1	-	-	-	-	-	15
4	1	23	10	-	-	-	-	-	34
7	5	14	27	-	-	1	-	-	47
10	3	7	28	6	1	2	-	-	47
13	11	9	5	18	1	-	-	-	44
16	3	37	6	18	8	-	-	-	72
19	3	27	22	10	16	-	-	-	78
22	-	8	29	6	16	11	-	-	70
25	1	-	21	16	4	24	2	-	68
28	-	1	9	28	4	8	10	-	60
31	-	-	1	10	7	2	16	-	36
34	-	-	-	7	10	3	-	5	25
37	-	-	-	-	6	4	2	2	14
40	-	-	-	-	1	1	-	-	2
43	-	-	-	-	-	-	1	-	1
46	-	-	-	-	-	-	1	-	1
Total	33	141	159	119	75	56	32	7	622

From the data of Tables II-IV, the sums, in units of grouping, are found to be :

$$S(x)=2,259 ; \quad S(y)=5,561 ; \quad S(z)=7,255$$

(x =difference in date of birth ; y =difference in age at test ; z =difference in age-adjusted score)

The means are, therefore :

$$\bar{x}=3.63183 ; \quad \bar{y}=8.94051 ; \quad \bar{z}=11.6640$$

Transforming these into months and points of score, the means are :

difference in date of birth=24.159 months ;

difference in age at test=18.822 months ;

difference in age-adjusted score=1.140 points.

The reason why the difference in age at test is smaller than the difference in date of birth is that the fourth age group was tested a year after the others. This also explains why in a few instances the older sib was younger at test than the younger sib.

The sums of squares and products, in units of grouping, are :

	<i>x</i>	<i>y</i>	<i>z</i>
<i>x</i>	9,839	22,221	26,400
<i>y</i>		55,781	65,045
<i>z</i>			91,223

and the sums of squares and products of deviations :

	<i>x</i>	<i>y</i>	<i>z</i>
<i>x</i>	1,634.69	2,024.38	51.05
<i>y</i>		6,062.80	181.57
<i>z</i>			6,600.75

The regression of *z* on *x* is given by : $51.05/1,634.69 = +0.03123$, and its standard error by :

$$\sqrt{\frac{6,600.75 - (0.03123 \times 51.05)}{620}} \times \frac{1}{1,634.69} = \pm 0.08069, \text{ etc.}$$

The regression coefficients of interest are :

$$z \text{ on } x = +0.03123 \pm 0.08069$$

$$z \text{ on } y = +0.02995 \pm 0.04189$$

$$y \text{ on } x = +1.23839 \pm 0.05923$$

$$x \text{ on } y = +0.33390 \pm 0.01597$$

or, in points of score per year :

$$z \text{ on } x = +0.7495 \pm 1.9366$$

$$z \text{ on } y = +1.1980 \pm 1.6756$$

The difference in age-corrected score between older and younger sibs is not significantly associated either with difference in date of birth or with difference in age at test. Nevertheless, it is the association with the difference in age at test which is the closer.

The sum of squares of deviations of difference in age at test at fixed difference in date of birth is :

$$6,062.80 - (1.23839 \times 2,024.38) = 3,555.83$$

and the sum of products of deviations—difference in score $\times$ difference in age at test

—at fixed difference in date of birth is :

$$181.57 - (0.03123 \times 2,024.38) = 118.35 \text{ etc.}$$

The sums of squares and products of deviations at fixed difference in date of birth are :

	<i>y</i>	<i>z</i>
<i>y</i>	3,555.83	118.35
<i>z</i>		6,599.16

and the sum of squares and products of deviations at fixed difference in age at test are :

	<i>x</i>	<i>z</i>
<i>x</i>	958.75	-9.58
<i>z</i>		6595.31

The partial regression of *z* on *y* at fixed *x* is :

$$118.35/3,555.83 = 0.03328$$

and its standard error :

$$\sqrt{\frac{6,599.16 - (0.03328 \times 118.35)}{619}} \times \frac{1}{3,555.83} = 0.05474$$

The two partial regression coefficients of interest are :

$$z \text{ on } y \text{ at fixed } x = +0.03328 \pm 0.05474$$

$$z \text{ on } x \text{ at fixed } y = -0.00999 \pm 0.10542$$

or, in points of score per year :

$$z \text{ on } y \text{ at fixed } x = +1.3312 \pm 2.1896$$

$$z \text{ on } x \text{ at fixed } y = -0.2398 \pm 2.5301$$

Thus such non-significant tendency as there is for difference in score to increase with difference in age is almost entirely accounted for by difference in age at test. This regression is almost unchanged when difference in date of birth is held constant, whereas when difference in age at test is held constant, the association between score and difference in date of birth is minus and practically zero.

The partial regression of difference in score on difference in age at test at fixed difference in date of birth may be used to eliminate the effect of difference in age at test. The superiority of the older sib was shown above to be +1.140 points.

$$+1.140 - (18.222/12 \times 1.3313) = -0.882$$

Any failure of the norms based on the whole sample to eliminate the association with age at test, as far as the 622 pairs of sibs are concerned, has not, therefore, minimized the non-significant advantage in favour of the older sib. On the other hand, the application of what seems an appropriate adjustment has reversed its sign.

IV. FAMILY SIZE, ORDER OF BIRTH AND INTELLIGENCE IN THE WHOLE GROUP

As described in the third paper of the series (1938a), number of living full sibs and test score was available for 3,305 children. The age-adjusted score used was an

Index of Brightness, calculated from the original norms and expressed as 100 plus or minus the deviation from the norm. An additional adjustment was incorporated to equalize the increase of variance of score with increasing age. The correlation between sib number and Advanced Otis I.B. was -0.224 , a value very similar to those obtained by several other investigators. My colleagues and I did not at that time analyse the data in relation to order of birth, though this had been recorded. As was mentioned in the introduction to the present paper, the outstanding difficulty is that some families are incomplete, the incompleteness itself being associated with the position of the child in the family. But while such material seems less suitable than the data of the previous section for investigating the possible influence of maternal age or of birth order, it does seem justifiable to use it for determining whether or not these factors contribute significantly to the association of family size and intelligence. This was done by Sutherland and Thomson (1926), who, using the method of partial correlation found that their simple correlation of -0.200 between birth order and intelligence was reduced practically to zero when family size was held constant. Although it is likely that any other method would yield substantially the same result it seems preferable in view of the non-normal distribution of family sizes and of birth orders to use analysis of variance and also to omit the larger families, which will be the most incomplete. Accordingly, I have taken families of up to five children only. A further comparison is made using only families of two and of three. The mean age of the child bringing the family into the survey was about eleven to twelve years at the time the information was obtained, and families which contained only two or three children at that time would probably be mostly complete.

TABLE V. FREQUENCY DISTRIBUTION OF OTIS I.B. BY FAMILY SIZE AND BIRTH ORDER

	Sibships of															Total
	Two		Three			Four				Five						
	1st	2nd	1st	2nd	3rd	1st	2nd	3rd	4th	1st	2nd	3rd	4th	5th		
Group central I.B.	4.5	1	—	—	3	—	2	—	—	4	—	—	—	—	1	11
	14.5	1	1	3	1	1	—	1	—	—	—	—	1	—	—	9
	24.5	1	3	3	—	4	—	4	—	3	—	—	2	1	2	23
	34.5	6	7	1	3	3	4	3	2	3	1	4	—	2	1	40
	44.5	9	7	8	6	9	1	8	4	7	1	2	—	3	2	67
	54.5	17	20	9	9	7	6	4	5	8	1	5	5	7	4	107
	64.5	14	25	8	11	12	5	11	8	6	5	4	3	3	6	121
	74.5	32	14	19	21	20	7	9	11	10	6	7	10	5	5	176
	84.5	31	29	27	25	18	9	12	14	17	2	7	3	8	11	213
	94.5	41	30	17	26	19	9	18	13	19	5	10	10	10	7	234
	104.5	56	57	16	25	27	12	17	20	13	1	3	8	7	6	268
	114.5	68	51	23	28	20	6	14	14	12	6	10	6	2	11	271
	124.5	35	54	21	24	16	6	13	14	13	2	5	8	2	4	217
	134.5	38	30	14	20	17	5	4	13	5	4	4	4	7	4	169
	144.5	24	37	15	11	18	7	5	4	6	2	—	—	2	5	136
	154.5	25	25	5	6	5	1	2	3	2	1	1	1	—	—	77
	164.5	8	8	7	2	2	3	4	3	1	—	2	1	—	3	44
	174.5	7	6	3	1	4	2	1	—	—	1	—	—	—	—	25
	184.5	2	3	1	1	1	—	1	1	—	—	1	—	—	—	11
194.5	—	—	—	1	—	—	—	3	—	—	—	—	—	1	5	
204.5	—	1	—	—	—	—	—	—	—	—	—	—	—	—	1	
Total	416	408	200	224	203	85	131	132	129	38	65	62	59	73	2,225	

The I.B.s based on the old norms have been used. As will be seen from Table XI, the difference from the new ones is small, and in this connexion there is no

reason to suppose that age at test is associated either with family size or with birth order. The data are presented in Table V. The numbers shown are slightly smaller than those previously given in Table III of the third paper of the series (1938a). This is because of the omission of a few idiots and imbeciles as already described, and because, unfortunately, the position in the family had not been filled in on a few of the cards.

The sums and means for family sizes and birth orders are given in Table VI, together with the necessary sums of squares of deviations.

TABLE VI. SUMS AND MEANS OF OTIS I.B. BY FAMILY SIZE AND BIRTH ORDER

Family size	Birth order	No.	Sum	Mean (units of I.B.)
	1	416	4,283	107.46
	2	408	4,272	109.21
Two		824	8,555	108.32
	1	200	1,952	102.10
	2	224	2,162	101.02
	3	203	1,964	101.25
Three		627	6,078	101.44
	1	85	796	98.15
	2	131	1,188	95.19
	3	132	1,316	104.20
	4	129	1,109	90.47
Four		477	4,409	96.93
	1	38	357	98.45
	2	65	581	93.88
	3	62	557	94.34
	4	59	497	88.74
	5	73	675	96.97
Five		297	2,667	94.30
Total		2,225	21,709	102.07

Sum of squares of deviations :

Whole sample .. 25,601.47

Families of 2 and 3 .. 16,206.58

The analysis of variance is based on the figures of Table VI and is shown in Table VII, using family sizes up to five.

TABLE VII. ANALYSIS OF VARIANCE. VARIATION OF I.B. IN RELATION TO FAMILY SIZE AND BIRTH ORDER. FAMILIES OF 2, 3, 4, AND 5

Variation of I.B.	Degrees of freedom	Sum of squares	Mean square	Variance ratio
Between family sizes	3	63,000	21,000.0	18.72
Between birth orders within family sizes	10	16,654	1,665.4	1.48
Within birth orders within family sizes	2,211	2,480,492	1,121.9	—
Total	2,224	2,560,146	—	—

The total sum of squares is 2,560,146 (after multiplying by 100 to convert from units of grouping to units of I.B.) and corresponds to 2,224 degrees of freedom. Perhaps the most convenient way of calculating the sum of squares between family size is :

$$\left\{ \frac{8,555^2}{824} + \frac{6,078^2}{627} + \frac{4,409^2}{477} + \frac{2,667^2}{297} \right\} - \frac{21,709^2}{2,225} \times 100 = 63,000$$

This corresponds to three degrees of freedom. If sums and means are used instead, the means must be calculated to eight or nine figures. The sum of squares between birth orders within family sizes is, for families of three :

$$\left\{ \frac{1,952^2}{200} + \frac{2,162^2}{224} + \frac{1,964^2}{203} \right\} - \frac{6,078^2}{627} \times 100 = 135$$

and similarly for other family sizes. The sum of squares within birth orders within family sizes is obtained by difference. The mean square between family sizes divided by the mean square within birth orders within family sizes gives a variance ratio of 18.72, which is highly significant. On the other hand the mean square between birth orders within family sizes when similarly divided gives a variance ratio of only 1.48. The tables show that for 12 and ∞ degrees of freedom the five per cent. point is 1.75, so the difference is not significant. Actually it lies fairly close to $P=0.2$.

An examination of the means of Table VI shows that the differences between the birth orders for families of two or three are remarkably small, as is brought out in the analysis of variance of Table VIII. The number in each birth order in families of four or five is small and the means considerably more variable. There seems, however, to be no suggestion of direction about this non-significant variation. In families of four the penultimate child is notably superior ; in families of five notably inferior. In families of four the last child is the dullest ; in families of five almost the brightest.

It is likely that families of two and three would mostly be complete, certainly more so than families of four or five. Accordingly, the analysis is repeated in Table VIII for these two sizes of family only. The difference between them is again highly significant and this time the mean square between birth orders within family sizes is substantially smaller than the mean square within birth orders within family sizes. Taking, therefore, those family sizes least likely to be affected by incompleteness there is no slightest suggestion of an association between birth order and intelligence.

TABLE VIII. ANALYSIS OF VARIANCE. VARIATION OF I.B. IN RELATION TO FAMILY SIZE AND BIRTH ORDER. FAMILIES OF TWO AND THREE

Variation of I.B.	Degrees of freedom	Sum of squares	Mean square	Variance ratio
Between family sizes	1	16,878	16,878.0	15.22
Between birth orders within family sizes	3	765	255.0	—
Within birth orders within family sizes	1,446	1,603,015	1,108.6	—
Total	1,450	1,620,658	—	—

The four family sizes selected for this test, and which include the bulk of the children, represent part of the regression of I.B. on family size, which was shown in the third paper of the series to be practically linear. Therefore it can be concluded that over this, the most important part of the range, birth order is not making any significant contribution to the association between intelligence and family size.

V. DISCUSSION

The results of this investigation have turned out to be very simple. They are in accordance with the conclusion of Jones (1933), whose careful review of the literature has already been mentioned. A sample of 622 pairs composed of an older and a younger child born to the same parents shows a small non-significant difference in favour of the older child. Such difference as there is is entirely due to a non-significant superiority of first-born over subsequent children. The advantage in favour of the older born child increases with increase in difference in date of birth, but it increases even more with difference in age at test, both associations being, however, non-significant. On partialling, it is found that increasing difference in date of birth is without effect when difference in age at test is held constant. Using this latter association to adjust slightly the age correction derived from the whole group from which the 622 pairs were drawn, the small non-significant difference in favour of the older born child is converted into a small and non-significant difference in favour of the younger.

The numbers are not large, however, and the mean difference in age between the sibs was only two years. All that can be said is that these results indicate that if advancing maternal age or increasing birth rank are unfavourable factors, their effect must be relatively small, at least when a sample is drawn from the whole population. This study could not, of course, throw much light on the possibility that there might be a more marked effect in the oldest maternal age groups.

The second part of the inquiry yields an equally definite result. The whole group of children gave a correlation of -0.224 between test intelligence and sib number. It is shown that no significant part of this association can be attributed to birth order. The fact that some of the families are incomplete raises difficulties; but when the smallest families only are analysed, that is, those most likely to be complete, the result remains the same. It is, perhaps, just conceivable that in some way incompleteness of families has reduced what would otherwise be a real contribution to non-significance, but this seems very unlikely. It appears reasonable to conclude that the association between the intelligence of children and the size of the families from which they come owes nothing significant to any presumed unfavourable effect of advancing maternal age or increasing birth rank.

VI. SUMMARY

1. The sample of 3,300 schoolchildren available was selected by age and geographical area only and tested on the Advanced Otis scale.

2. There were included the equivalent of 622 pairs of sibs born to the same parents and separated by an average of two years. Age correction is based on norms calculated from the whole sample. There is a small and non-significant mean difference in age-corrected score in favour of the older member of the pair. Division of the pairs, first, into those containing and those not containing a first-born child, and secondly, by sexes, reveals no significant difference.

3. When age at test is held constant, there is no increase in the difference between older and younger sibs with increasing interval between dates of birth. At fixed difference in date of birth and zero difference in age at test (i.e., adjusting the norms to the 622 pairs) the small non-significant difference in favour of the older born sib is changed to a small non-significant difference in favour of the younger.

4. It is concluded that any effect of advancing maternal age or advancing birth rank on intelligence must be relatively small.

5. Examining families of two, three, four, and five in the whole sample by

family size and birth order, it is found that differences in intelligence between family sizes are highly significant and differences between birth orders within family sizes non-significant. This is equally true for families of two and three only, i.e., those most likely to be complete. It is concluded that birth order, and hence maternal age, make no significant contribution to the correlation of -0.224 between sib number and Otis I.B. found in the whole sample.

6. The norms previously used for age correction were slightly biased by the order in which schools were visited. An attempt to eliminate this bias is described in the Appendix.

ACKNOWLEDGEMENTS

I wish to express my indebtedness to Dr. R. M. Norman and Dr. Ruth Griffiths, who collaborated in the original survey, and as well to all those in the City of Bath whose help made the survey possible. I am also much indebted to Dr. J. O. Irwin for advice in connexion with appropriate methods for securing unbiased estimates for age correction, as described in the Appendix, and on the test of significance employed in Table VII, and to Dr. W. J. Martin for advice in regard to the presentation of some of the tables.

REFERENCES

1. C. Burt (1947) *Intelligence and Fertility*. London: Hamish Hamilton.
2. E. A. Cockayne and L. S. Penrose: 'The genetics of congenital pyloric stenosis.' *Ohio Journ. of Science*, XLIII, 43.
3. H. E. Jones (1933): 'Order of birth.' *Handbook of Child Psychology*, 2nd ed., Worcester, Massachusetts.
4. P. Malpas (1937): 'The incidence of human malformations and the significance of changes in the maternal environment in their causation.' *Journ. of Obstetrics and Gynaecology*, XLIV, 434.
5. L. S. Penrose (1934): 'The relative aetiological importance of birth order and maternal age in mongolism.' *Proc. Roy. Soc. B*, CXV, 431.
6. — (1938): 'Some genetical problems in mental deficiency.' *Journ. of Mental Science*.
7. J. A. F. Roberts, R. M. Norman, and R. Griffiths (1935): 'Studies on a child population. I. Definition of the sample, method of ascertainment, and analysis of the results of a group intelligence test.' *Ann. Eugen.*, VI, 319–38.
8. J. A. F. Roberts and R. Griffiths (1937): 'Studies on a child population. II. Retests on the advanced Otis and Stanford-Binet scales with notes on the use of a shortened Binet scale.' *Ann. Eugen.*, VIII, 15–45.
9. J. A. F. Roberts, R. M. Norman, and R. Griffiths (1938a): 'Studies on a child population. III. Intelligence and family size.' *Ann. Eugen.*, VIII, 178–215.
10. — (1938b): 'Studies on a child population. IV. The form of the lower end of the frequency distribution of Stanford-Binet intelligence quotients and the fall of low intelligence quotients with advancing age.' *Ann. Eugen.*, VIII, 319–36.
11. J. A. F. Roberts (1940): 'Studies on a child population. V. The resemblance in intelligence between sibs.' *Ann. Eugen.*, X, 293–312.
12. — (1945): 'On the difference between the sexes in dispersion of intelligence.' *Brit. Med. Journ.*, I, 727.
13. H. E. G. Sutherland and G. H. Thomson (1926): 'The correlation between intelligence and size of family.' *Brit. Journ. of Psych.*, XVII, 81.

APPENDIX

The removal of sources of bias in the calculation of norms for the Advanced Otis Scale (Form A)

The whole sample of schoolchildren at Bath consisted of four complete age groups. The testing of the youngest of these was postponed for a year as they were

too young for the test. Accordingly the means and regressions of score on age were based on the first three age groups. At the private schools (except one) the testing of all four age groups was carried out at a time intermediate between the two series of visits to the other schools. From the private school fraction children were included in the calculations according to date of birth ; in this way the private schools yielded a sample corresponding in age at test to the three oldest age groups at the other schools. It is well known that scales of this type yield a regression of score on age which is substantially linear up to the age of twelve. Thereafter the rate of increase falls off sharply. This proved true in our material, so two regression lines were fitted, one up to and including 12.0 years, one down to and including 12.0 years. Neither showed any significant departure from linearity. We followed Otis in calculating an Index of Brightness (I.B.), which is simply 100 plus or minus the deviation from the expected score. We incorporated, however, a correction which equalized differences from the norm at different chronological ages, for variance of score increases rapidly with advancing age. The calculations were described in the first paper of the series (1935).

We soon became aware, however, that the estimates of the regression coefficients were not strictly unbiased. This was due to the fact that the testing, which was all carried out by three of us, took a considerable time. The sample was defined by date of birth and so the children were growing away from us during the period of the testing. Even amongst junior elementary schools the mean difference in performance is relatively large. This means that if there is a tendency for the schools containing the cleverer children to be visited first and those containing the duller children last, the regression will be flattened, for the duller children will fall into age groups, say for example, two months older than the cleverer children. Randomization of the order in which schools were visited would not have helped, as the number of schools was, of course, relatively small, and any order chosen, while it might be random, would still in all probability mean that duller children were tested first or *vice versa*. In any event, this was not a practical procedure. The testing of the first three age groups, apart from absentees and private school children, occupied about five months, the junior elementary schools being visited first. These contained all the third age group and most of the second. We spent the first month at the largest school, containing about 15 per cent. of the entire sample under 12.0 years, as we were experimenting before committing ourselves to our final procedure ; this school was well above the average in the intelligence of its pupils, as was the second school visited (though not the third). In retrospect there is no doubt that there was a tendency for schools containing the brighter children to be visited earlier. This portion of the regression was therefore artificially flattened to some extent. Secondary, central, and senior elementary schools show, of course, very wide differences in mean brightness. There were, however, only nine of them, and all the testing was completed within a month, so that any bias could only have been small.

The second source of bias arises from premature promotion in the second age group. A proportion of the very brightest had gone on to the central and secondary schools a year before they need have done. Those who had moved on to senior elementary schools were also well above the average. As the junior schools were visited first, in this instance it was the duller children who were tested first and the relatively small but very bright fraction later. Although the previous regression up to 12.0 was not significantly non-linear, indications of the effect of the two sources of bias can be seen in the diagram at page 330 of the first paper (1935). The lower part on the curve is unduly flattened and the upper part unduly steep.

It was indeed a fortunate circumstance that the two sources of bias operated in opposite directions, the result being that the new norms of this paper do not differ

seriously from the old. The use of the old I.B.s cannot have affected appreciably previous comparisons made on this sample of children, for example by sex, or by season of birth (as there were four age groups). But the comparison of the 622 pairs of older and younger sibs made in the present paper would have been impossible without the elimination of these sources of bias, small though they proved to be. What is required is the division of the material into units followed by analysis of variance and co-variance. The unit is the school-visit, comprising the few days, at most, during which the testing was carried out at each school. The second age group, with its proportion of prematurely promoted children, required special treatment. Apart from private schools and absentees, the division into groups was carried out as follows for children up to age 12·0 :

First age group (September, 1921–August, 1922)

Some of the younger members of this group were 12·0 or younger at test. They were at senior, etc., schools and so automatically fall into separate school-visit units.

Second age group (September, 1922–August, 1923)

The prematurely promoted children were practically all born during the months September–December, 1922. Hence the remainder, practically all being at the junior schools, could be added to the third age group tested at the same time and place. The September–December children had to be treated as separate units involving schools of all types.

Third age group (September, 1923–August, 1924)

All at junior schools and added to January–August children of second age group.

Fourth age group (September, 1924–August, 1925)

Tested a year later and so kept separate from the remainder.

Apart from two schools taking special place (scholarship) children from the elementary schools, and which therefore required special treatment, the question of promotion and of the postponement of the test did not arise at the private schools. Each school therefore could be treated as a single unit.

At the large elementary school already mentioned, absentees were tested before the end of the summer term of 1934. They form, therefore, a separate unit. At the other schools, absentees (apart from single children given a Binet test only) were tested during a relatively short period with well mixed up order of visit to schools of all types, so it seems justifiable to add them together to form a single unit. Altogether these school visits, with the further sub-division of the second age group, numbered exactly 100. It is clear that the bias in the estimate of the regression should lie between units, and that the regression within units should provide the unbiased estimate required for calculating the norms.

The results of the analysis are shown in Table IX.

TABLE IX. CHILDREN UNDER 12 YEARS 1 MONTH. ANALYSIS OF VARIANCE AND CO-VARIANCE. AGE AT TEST AND SCORE

Variation				Degrees of freedom	Sum of squares		Sum of products
					Age ²	Score ²	Age × Score
Between units	..	..	..	99	94,794	7,694	13,023
Within units	..	..	..	2,371	59,833	23,222	10,104
Total	..	..	..	2,470	154,627	30,916	23,127

The sums of squares and products of deviations were calculated separately for each of the 100 units. The sums of the sums give the sums of squares and deviations within units shown in Table IX. The totals for the complete group treated as a single unit are shown in the last line. The sums of squares and products between units are obtained by subtraction.

The regressions on age at test therefore are :

Between units	0.13738 ± 0.02534
Within units	0.16887 ± 0.01232
Total	0.14956 ± 0.00848

These are in units of grouping of one month and ten points of score. Converting into points of score per year, they become :

Between units	16.486 ± 3.041
Within units	20.264 ± 1.478
Total	17.947 ± 1.018

The differences are not significant, but the regression within units should be the unbiased estimate required. The figure used in calculating the old norms was :

18.344

The treatment of the group of age 12.0+ was simpler as all had passed the promotion age. The analysis is shown in Table X.

TABLE X. CHILDREN OVER 11 YEARS 11 MONTHS. ANALYSIS OF VARIANCE AND CO-VARIANCE. AGE AT TEST AND SCORE

Variation	Degrees of freedom	Sum of squares Age ²	Sum of squares Score ²	Sum of products Age $\times$ Score
Between units	36	2,457	5,142	315.4
Within units	861	10,602	7,250	558.1
Total	897	13,059	12,392	873.5

Regressions

In units of grouping :	Between units	=	0.12837 ± 0.24712
	Within units	=	0.05264 ± 0.02816
	Total	=	0.06689 ± 0.03248

In points per year :	Between units	=	15.404 ± 29.654
	Within units	=	6.317 ± 3.379
	Total	=	8.026 ± 3.898

Regression used for old norms = 5.120 points per year.

The very large difference between the regressions between and within units is due to the different time of visiting private schools and had been allowed for, as already explained, when the old norms were calculated.

The figures required for the calculation of a table of norms are :

Children below 12 years 1 month

Mean age=10 years 8·1076 months.

Mean score=72·9379 points.

Regression of score on age (within units)=20·264 points per year.

Children above 11 years 11 months

Mean age=12 years 5·6024 months.

Mean score=102·9744 points.

Regression of score on age (within units)=6·317 points per year.

Table XI gives the new norms, with the old ones added in brackets.

TABLE XI. NORMS FOR ADVANCED OTIS SCALE, FORM A. PREVIOUS NORMS IN BRACKETS

Months				9 years	10 years	11 years	12 years	13 years
0	..	..	..	—	59 (61)	80 (79)	100 (98)	106 (105)
1	..	..	..	—	61 (63)	81 (81)	101 (99)	107 (105)
2	..	..	..	—	63 (64)	83 (82)	101 (100)	107 (106)
3	..	..	..	—	64 (66)	85 (84)	102 (101)	108 (106)
4	..	..	..	—	66 (67)	86 (85)	102 (101)	108 (106)
5	..	..	..	47 (50)	68 (69)	88 (87)	103 (102)	109 (107)
6	..	..	..	49 (52)	69 (70)	90 (89)	103 (102)	—
7	..	..	..	51 (53)	71 (72)	91 (90)	104 (103)	—
8	..	..	..	52 (55)	73 (73)	93 (92)	104 (103)	—
9	..	..	..	54 (56)	74 (75)	95 (93)	105 (103)	—
10	..	..	..	56 (58)	76 (76)	96 (95)	105 (104)	—
11	..	..	..	58 (59)	78 (78)	98 (96)	106 (104)	—

Values for ages outside the range 9 years 10 months to 12 years 10 months are extrapolations to regions with few observations.

A study of Tables IX and X, with the regressions, reveals the value of planning in a task of this type. In the younger age group a good deal of precision in estimation has been lost, due of course to the narrow ranges of age at test within the separate units, which reduces the denominator in calculating the standard error. This is partly compensated for, however, by the lower variability of score within the units, which reduces the numerator. In those over 12·0 the standard error of the regression within units is actually smaller than that of the total. Quite apart from the difficulties of selecting a good sample of schools and of children with respect to intelligence, there must always remain difficulties to be overcome in arranging the testing unless the test is of a self-administering type and can be applied simultaneously to the entire group. Selection of children by chronological age at date of visit would go some way to eliminate bias, but the existence of age groups which are partly promoted must always remain a troublesome factor. An appropriate division between and within units can increase the accuracy of estimation of the required regressions provided that the units can be chosen so as to retain as much variability as possible in age and as little as possible in intelligence.

THE VARIATIONS OF INTELLIGENCE WITH OCCUPATION, AGE, AND LOCALITY

By P. E. VERNON

Psychology Department, University of Glasgow

I. *Introduction.* II. *The Consistency of the Occupational Classification.* III. *The Separation of the Three Factors.* IV. *Results.* V. *Discussion.* VI. *Summary.*

I. INTRODUCTION

During four years of the war, the Progressive Matrices Test (Raven, 1939) was applied, with a 20-minute time limit, by specially trained Wren Petty Officer Recruiting Assistants to all candidates for the Royal Navy at Combined Recruiting Centres. Careful records were kept of scores, ages, occupations, and other data both for accepted and for rejected men. This article analyses some of the material collected from 89,764 men tested during part of 1943. When the pressure of work fell off somewhat in 1944-5, the Wrens were asked to extract from the records and to tabulate on specially designed forms the scores and ages of men falling within twelve main occupational groups (shown in Table I). Roughly a quarter of the candidates who fell under none of these headings were to be excluded. Each Wren dealt with men tested in the particular Centre or Centres where she worked, but it should be

TABLE I. OCCUPATIONAL GROUPS AND FREQUENCIES

Occupational Group	Number
1. Clerks, all types ; stock-keepers, book-keepers	11,603
2. Drivers, all except horse ; tractor and crane drivers	5,764
3. Builders ; bricklayers, plasterers, painters, slaters, apprentices, <i>not</i> mates ..	7,074
4. Precision Workers ; fitters, turners, centre lathe operators, toolmakers, <i>not</i> setters	11,067
5. Sheet Metal Workers ; plumbers, welders	6,401
6. Machine operators ; machinists, capstan operators, grinders, millers, setter-operators	5,936
7. Woodworkers ; joiners, carpenters, cabinet makers	6,705
8. Electrical Workers ; wiremen, linesmen, electrical and GPO engineers, radio servicemen	4,602
9. Retail Tradesmen ; shop assistants, roundsmen, bar tenders, <i>not</i> travellers ..	9,373
10. Farm Workers, all except tractor drivers	2,406
11. Labourers, all except farm	13,466
12. Mates	5,367
Total	89,764

noted that the majority of Wrens would not themselves have applied the test to, nor written out the records of, the candidates, since they would either have been promoted, or have moved to other Centres or jobs between 1943 and 1944-5.

The age groups chosen and numbers in each are given in Table II.

TABLE II. AGE GROUPS AND FREQUENCIES

Age group	Number
1. 16-17 yrs., the great majority being 17	32,412
2. 18-19 yrs., 19-year men were tabulated separately, but being relatively few were later combined	30,451
3. 20-24 yrs.	6,165
4. 25-29 yrs.	4,060
5. 30-34 yrs.	5,351
6. 35-39 yrs.	7,577
7. 40 yrs. and over	3,748
Total	89,764

The eighty odd Recruiting Centres were combined under the nine Naval Recruiting Areas shown in Table III. From the anthropometric standpoint these areas are singularly ill-chosen. For example, Glasgow includes the Glasgow Irish and the Scottish Highlanders. Rural and urban are mixed in all areas, though rural presumably predominates in Southampton.

TABLE III. AREAS AND FREQUENCIES

Areas	Number
1. Birmingham, including Gloucester, Hereford, Worcester, Shropshire and mid-Wales	9,397
2. Bristol, including Wilts, Somerset, Devon, Cornwall and S. Wales	7,988
3. Derby, including Lincoln and Nottingham	5,496
4. Glasgow, including the whole of Scotland	11,223
5. Liverpool, including W. Lancashire and N. Wales	7,690
6. London, including Kent, E. Anglia and Northampton	22,376
7. Manchester, including Yorkshire	12,625
8. Newcastle, including Cumberland, Westmorland and Durham	5,264
9. Southampton, including Sussex, Oxford and Dorset	7,705
Total	89,764

The tabulations were recorded on Hollerith cards and, in 1946, score distributions were sorted by Occupation, Age, and Area ; i.e., there were $12 \times 7 \times 9 = 756$ distributions. Scores were classified into groups of 5 (55+, 50+ . . . 10+, 5+) and totalled by hand.¹

II. THE CONSISTENCY OF THE OCCUPATIONAL CLASSIFICATION

A certain subjective element enters into any occupational classification, and as a large number of Wrens undertook the tabulation it is desirable to study their consistency. This was done in a separate experiment by taking two specimen record books, containing 636 cases, and having them classified independently by 14 Wrens. Two Wrens made such serious errors in interpreting the instructions that it was judged best to exclude them. Numbers in each occupational group, and in the unclassified group, were averaged for the other twelve, and these averages were used as expected frequencies. Values of chi-squared for separate occupations and Wrens are shown in Table IV.

¹ Only one man scored 60/60 ; he was grouped with the 55+ men. No one scored less than 5.

TABLE IV. VALUES OF CHI-SQUARED SHOWING INCONSISTENCIES AMONG WRENS IN CLASSIFYING OCCUPATIONS

Occupation					Chi-Squared		Wren No.		Chi-Squared
7.	Woodworkers	..	..	..	0.76		1	..	2.77
12.	Mates	..	..	..	1.39		2	..	2.86
2.	Drivers	..	..	..	1.50		3	..	4.46
3.	Builders	..	..	..	3.39		4	..	4.65
1.	Clerks	..	..	..	3.43		5	..	7.09
9.	Tradesmen	..	..	..	5.79		6	..	7.62
5.	Sheet Metal Workers	..	..	..	8.56		7	..	8.16
4.	Precision Workers	..	..	..	8.68		8	..	8.37
8.	Electricians	..	..	..	9.20		9	..	9.72
10.	Farm Workers	..	..	..	13.93		10	..	9.98
6.	Operators	..	..	..	21.01		11	..	10.22
11.	Labourers	..	..	..	24.81		12	..	26.55
Total ..					102.45		Total	..	102.45
Unclassified					77.62				

By far the greatest discrepancies occur among the cases left unclassified. The mean Wren excluded 168 cases, but the range for different Wrens was from 91 to 218; and when all 14 Wrens were considered the range was from 2 to 273. However, we are only concerned with classified cases, and here the consistency is quite high. Assuming 11 degrees of freedom, only two occupations—Operators and Labourers—show a departure from expectation with $P < .05$, and in no occupation is $P < .01$. Again of the Wrens, only No. 12 shows a significant departure from expectation so long as unclassified cases are omitted. Most Wrens exhibited a few minor idiosyncrasies; for instance, No. 1 had 28 instead of 35 Farm Workers, No. 5 had 32 instead of 25 Electricians, and so on. It should be pointed out, however, that the Wrens in the Recruiting Centres who undertook the main tabulation would usually consult with their colleagues when in any doubt; whereas in this special experiment the 14 Wrens worked independently throughout and so might be more liable to error. Thus although it is possible that a few of the tabulators made serious mistakes, and although the Operator and Labourer categories are rather poorly defined, yet we may conclude that the occupational totals for any one age or area group are likely to be quite reliable when the tabulations for all Recruiting Centres are combined.

III. THE SEPARATION OF THE THREE FACTORS

The obtained differences between occupational, age, and area groups are misleading because of the overlapping between these three factors. Builders, for example, come out very low because they happen to contain the largest proportions of older men who were de-reserved at that time, and Mates are too high because they are predominantly younger men (cf. Table V). Conversely, age differences may be affected by the much larger proportion of Clerks than of Builders in the younger groups. Glasgow area is low again partly because it included a very big proportion of Labourers, and Manchester high because it had an excess of Clerks and Electricians and few Farmers, Labourers or Mates (cf. Table VII).

It was not possible to separate the effects of these factors directly by analysis of variance because of the very uneven numbers in the 756 distributions. For example, there were 1,988 Clerks in London aged 16–17, and only one Farm Worker and one Mate in Newcastle aged 40+. Two courses appeared to be open:

(i) Independence values, or numbers to be expected in each distribution were

obtained by a kind of three-dimensional chi-squared procedure. Thus we should expect :

$$\frac{11,603 \times 22,376 \times 32,412}{89,764^2} = 1,044.37 \text{ Clerks in London aged 17,}$$

$$\text{and } \frac{2,406 \times 5,264 \times 3,748}{89,764^2} = 5.89 \text{ Farm Workers in Newcastle aged 40+}.$$

The obtained numbers in each distribution were then reduced or increased accordingly, and the mean scores added up afresh. Although this plan was actually followed in the first analysis of variance quoted below, it is far from satisfactory. Clearly it is unnatural to separate Occupation and Area. We would expect there to be more Farm Workers in Southampton than in Liverpool, and more Precision Workers in Liverpool than Southampton, hence the mean scores for Areas are falsified when all Areas are made to contain the same proportions of Farm and Precision Workers.

Age, again, possesses a slight natural overlapping with Occupation ; for instance, few men stay as Operators or Mates to a late age. But we shall not greatly distort the facts if we assume the age distribution to be constant in all occupations. Any overlapping between Age and Area is obviously fortuitous and would be better eliminated.

(ii) An alternative plan, therefore, was to retain the original Occupation $\times$ Area totals, and to calculate independence values for Age groups. Thus there were actually 3,579 Clerks in London, and their expected age distribution was found from the general age distribution given in Table II. This yields :

$$\frac{3,579 \times 32,412}{89,764} = 1,292.3 \text{ London Clerks aged 17.}$$

Similarly there should be 4.80 Farm Workers and 8.56 Mates aged 40+ in Newcastle.

With both plans the problem then arose of how to make up the deficits in certain cells. If only one Farm Worker and one Mate were tested in Newcastle, how could scores for 4.80 and 8.56 respectively be estimated ? It was decided to reduce the total number

of cases to 10,000. Then it was only necessary to have : $\frac{1,292.3 \times 10,000}{89,764} = 144$

London 17-year Clerks, and similarly 0.54 and 0.95 Newcastle 40+ Farmers and Mates. The expected frequencies were all taken to the nearest whole number, so one Farmer and one Mate were needed. Actually all but 2 per cent. of the cells could now be filled, and the slight deficits were made up by borrowing excess cases of the same Occupation and Age from other Areas.

From the point of view of the subsequent analysis of variance, it would have been far better to select the 10,000 cases at random from the 89,764 till all the cells were filled. But it is difficult to think of any good method of picking at random from frequency distributions ; and it would have meant retabulating 10,000 scores. Instead therefore the obtained distributions were reduced ; for example, the 1,988 London 17-year Clerks' distribution was reduced to 144. This meant, of course, that the distributions for those cells which underwent big reductions were far more reliable than those for other cells where there were only just sufficient (or insufficient) cases to make up the expected numbers. Thus it will be difficult to make true estimates of the significance of the several sources of variance. The total Occupation, Age, and Area distributions will be considerably more reliable than if they had been based on 10,000 unselected cases ; on the other hand the interactions, particularly the second order one, may be too small. Probably a better estimate of the effective total number of cases would be about 30,000. However, this is the kind of difficulty that always arises when the psychologist attempts to analyse data collected in the ordinary course of vocational or educational practice, when, in other words, he is unable to impose a desired experimental design.

In reducing the distributions, frequencies were carried to the first decimal place. For example, there were originally 10 London 40+ Farm Workers, but only two were needed to fill the cell. The original distribution was therefore converted as follows :

Score		Obtained Frequency		Adjusted Frequency
45+	..	2	..	0.4
40+	..	1	..	0.2
35+	..	4	..	0.8
30+	..	1	..	0.2
25+	..	1	..	0.2
20+	..	1	..	0.2
		<u>10</u>		<u>2.0</u>

These adjusted distributions were then summed to give the Occupation, Age, and Area distributions listed in the Appendix.

IV. RESULTS

Means. Three sets of mean scores are quoted in Tables V-VII. The first column in each Table is based on the original 89,764 scores ; the second gives the means when Occupation, Age, and Area are all made independent, as in our first plan. In the third column, Occupation and Area are made independent of Age, but not of each other, as in our second plan, and these figures are regarded as the most representative. The actual distributions from which this third set of means was obtained are given in the Appendix.

TABLE V. MEAN SCORES FOR OCCUPATIONS

Occupation	(1)	(2)	(3)
1. Clerks	42.96	42.29	42.43
8. Electrical Workers	39.54	39.25	39.20
4. Precision Workers	38.24	37.81	37.76
7. Woodworkers	36.19	37.33	37.35
5. Sheet Metal Workers	36.11	36.58	36.56
6. Operators	36.37	35.63	35.58
9. Retail Tradesmen	36.58	36.13	35.27
3. Builders	33.01	35.10	35.07
12. Mates	35.84	33.96	34.24
2. Drivers	33.61	33.70	33.63
10. Farm Workers	32.73	32.61	32.33
11. Labourers	32.43	32.04	31.86
Average	36.45	36.31	36.29

TABLE VI. MEAN SCORES FOR AGE GROUPS

Age	(1)	(2)	(3)
16-17	37.97	37.70	37.68
18-19	36.94	37.15	37.13
20-24	36.86	36.43	36.43
25-29	35.28	34.69	34.66
30-34	34.49	33.73	33.74
35-39	32.82	32.52	32.40
40+	30.33	30.34	30.30

TABLE VII. MEAN SCORES FOR AREAS

Areas	(1)	(2)	(3)
London	37.48	37.26	37.51
Manchester	37.35	36.73	37.24
Derby	36.57	36.52	36.51
Southampton	36.72	36.76	36.26
Newcastle	36.36	36.17	35.96
Liverpool	35.99	36.02	35.81
Birmingham	36.00	35.46	35.53
Bristol	34.89	35.19	34.96
Glasgow	34.83	35.29	34.78

Occupations and Areas are listed in order of the third set of means. These orders conform closely to expectation, the most intellectual and highly skilled occupations being top, together with London area, and Bristol and Glasgow areas (with their large Welsh and Irish components) lowest. It should be remembered that the Matrix test is a non-verbal one, as free as possible from educational influence; thus the differences are much smaller than those obtained with some verbal tests such as Army Alpha or Cattell Scale III. Clerks are only 0.71σ above, and Labourers 0.52σ below the general mean, where $\sigma = 8.58$. The Age differences, on the other hand, are more marked than with verbal tests, and the second and third columns show almost perfectly linear regression. But this result should be accepted with caution for two reasons. First the 16-17-year-olds are volunteers who are known to score on the average two to three points higher than conscripts, whereas the 18-year and later groups contain unknown but certainly smaller proportions of volunteers. Secondly, the groups of 20 years and over consist mainly of men who were in reserved occupations which had recently been dereserved, hence they are much less representative of the population as a whole than the 18-year conscripts.

Area differences are small, but it has already been pointed out that our Areas are too mixed to bring out the full range of geographical variations.

The first column of Occupational means is badly distorted by the irregular age distributions. As mentioned above, Mates and Operators come out unduly high, and Builders too low. But the second and third columns are almost identical. In the Table for Areas these two columns are more discrepant. Thus Manchester's mean drops in the second column when its proportion of skilled workers is reduced to that in the country as a whole. Similarly Glasgow's mean rises when its undue number of Labourers is allowed for, and it slightly surpasses the Bristol area.

Analysis of Variance. Analyses based on the two plans are shown in Table VIII.¹ The contributions of Occupation, Age, and Area are 12.83%, 5.65%, and 0.81% in the first, and 13.44%, 5.67%, and 1.40% in the second analysis. Age remains almost the same in the two analyses, but Occupation and Area overlap so much that their Interaction term becomes a minus quantity in the second plan. In other words Occupation's 3,955.39 includes a good deal due to Area, while Area's 412.36 includes a still larger proportion due to Occupation. Area's contribution, once Occupation is held constant, is very small (though still statistically significant).

A third alternative is to calculate Variance between Occupations within Areas, and Between Areas within Occupations. These amount to 3,849.91 and 306.88 and

¹ In the first plan total variance was not recalculated, but assumed to be the same as in the second. The total variance for the original distributions (reduced to 10,000 cases) was 28,609.81; the second plan's variance of 29,421.65 was so close that an additional conversion and summation of the 756 distributions seemed unnecessary.

TABLE VIII. ANALYSES OF VARIANCE

Source of Variance	(First Plan)			(Second Plan)	
	Sums of Squares	Degrees of Freedom	Mean Square	Sums of Squares	Mean Square
Between Occupations	3,775.96	11	342.27	3,955.39	359.58
Between Ages	1,662.93	6	277.15	1,667.14	277.86
Between Areas	237.92	8	29.74	412.36	51.54
Interactions—					
Occ. × Age	196.63	66	2.98	197.95	3.00
Occ. × Area	77.28	88	0.88	(-105.48)	1.20
Age × Area	43.32	48	0.90	41.29	0.86
Occ. × Age × Area	403.79	528	0.76	374.01	0.71
Individual Variance	23,023.82	9,244	2.49	22,878.99	2.48
Total	29,421.65	9,999		29,421.65	

their Interaction to +105.48. Unfortunately no exact technique could be discovered for assessing the separate and the joint contributions of Occupation and Area, as well as the Interaction. But presumably the differences between the two plans, namely $3,955.39 - 3,775.96 = 179.43$, and $412.36 - 237.92 = 174.44$, provide a good guess as to the variance of the inter-correlated effect. Thus the following balance sheet appears to give the best picture of the make-up of Matrix scores :

	Variance per cent.
Individual differences	77.66
Occupational differences	12.83
Area differences	0.81
Combined effect due to overlap of Occupation with Area	0.60
Regression of Age ¹	5.65
Differential regression of Age in different Occupations ¹	0.67
Irregularities (other interactions)	1.78
	<u>100.00</u>

This brings out well the truth of the statement often made by psychologists that individual differences within occupational groups are much larger than differences between these groups. It shows, too, the relative unimportance of all factors other than Occupation and Age covered by this investigation.

Error Variance. It was suggested above that the second order interaction would under-estimate, and individual variance over-estimate, the error. This is borne out by our figures, and if we accepted 2.49 as the Mean Square discrepancy the interactions would be significantly too small. But the Occupation × Area and Age × Area first order interactions in the first plan yield almost identical estimates, 0.88 and 0.90, and their mean, 0.89, would seem to be the fairest estimate of error. Had the total population been reduced to 28,000 instead of 10,000, this .89 would have risen to 2.49, and the F-ratios of these interactions would be 1.0. The second order interaction would still be too small, but its F-ratio of 1.17 would be doubtfully significant even with 528 degrees of freedom.

¹ These two terms are explained in a later section.

Actually the variances of Occupation, Age, and Area are so highly significant that the size of the Error Variance hardly matters. It is important, however, in relation to the Occupation $\times$ Age interaction, and if we accept 0.89, the F-ratio of 3.35 is fully significant with 66 d.f. This must mean that scores do not fall uniformly with age in different occupations.

Effects of Age. As indicated above, it is doubtful how far the obtained effect is purely due to age, how far to differences in proportions of volunteers or of dereserved men at different ages. Nevertheless, the differential effect calls for further analysis. A plausible hypothesis is that the decline with age is greater in the lower occupational groups. Cattell (1943) claims to have found this, and Miles's (1932) investigation (on very small numbers) reveals an earlier decrease in intelligence among the most poorly educated adults.

In our group as a whole the correlation ratio for scores on Age is .238 but ratios for different occupations (shown in Table IX) range from .188 and .193 among Tradesmen and Woodworkers to .325 and .373 among Labourers and Mates. An alternative index of the extent of decline, which is less affected by the uneven numbers and chance fluctuations in age-group means, is the difference between the scores at 16-19 and at 30-40+. This index ranges from 3.40 and 3.72 in Woodworkers and Sheet Metal Workers to 6.99 and 8.49 in Labourers and Mates. It yields a rank order correlation of -0.44 with the mean scores of all the occupational groups. Still more striking are the figures, given in Table IX, for the two largest samples—the 16-17

TABLE IX. CORRELATION RATIOS, AND DECREASES BETWEEN 16-19 AND 30-40+ AND BETWEEN 16-17 AND 18-19, IN DIFFERENT OCCUPATIONS

Occupation	Eta	Difference 16+ to 30+	16-17 yrs.		18-19 yrs.		Decrease
			No.	Mean	No.	Mean	
Clerks	.273	5.07	6,226	43.64	3,889	43.25	0.39
Electrical Workers	.277	5.18	1,266	40.34	1,754	40.18	0.16
Precision Workers	.280	5.43	2,723	39.06	4,416	38.73	0.33
Woodworkers	.193	3.40	1,067	38.09	1,174	37.84	0.25
Sheet Metalworkers	.203	3.72	1,404	37.66	1,698	37.17	0.49
Retail Tradesmen	.188	3.80	4,865	37.28	3,372	36.51	0.77
Machine Operators	.272	5.67	2,547	36.91	2,131	36.49	0.42
Mates	.373	8.49	3,031	36.55	1,955	35.70	0.85
Builders	.224	4.43	1,162	36.25	1,096	35.48	0.77
Drivers	.234	4.66	1,498	34.87	2,618	34.36	0.51
Labourers	.325	6.99	5,935	34.38	5,026	32.81	1.57
Farm Workers	.251	5.17	688	34.07	1,322	32.68	1.39
All occupations	.238	5.26	32,412	37.81	30,451	37.09	0.72

and 18-19-year-olds. Even over this small age range the decrease in score is smallest in the most intelligent occupations, averaging less than 0.3 among Clerks, Electricians, Precision and Woodworkers, and is five times as great in the least intelligent Labourers and Farmworkers. A further analysis of variance of these figures (Table X) shows that the differential effect, represented by the Interaction term, is highly significant. The rank correlation between intelligence level and decrease is -0.86.

TABLE X. ANALYSIS OF VARIANCE BETWEEN 16-17 AND 18-19 YEARS AND BETWEEN OCCUPATIONS

Source of Variance	Sums of Squares	Degrees of Freedom	Mean Square	F-ratio
Between Occupations ..	26,003.371	11	2,363.943	5.401 (P= <.001)
Between Ages	326.023	1	326.023	
Interaction	133.512	11	12.137	
Individual Variance ..	141,214.113	62,839	2.247	
Total	167,677.019	62,862		

This result may perhaps be distorted by the volunteer-conscript factor, for there does appear to be a slightly smaller proportion of volunteers among the lower occupational 18-19 year groups. A fresh tabulation was made for 5,207 men of this age, and the proportions were 20.8% in the four highest occupations, 18.3% in the middle three (Sheet Metal Worker, Machine Operator, Driver), and 15.8% in the five groups showing the biggest decreases. But it was calculated that this would only affect the score-decrease by a small amount; it might make it 0.13 larger in the lower than the upper groups, whereas our obtained difference is 0.85 larger. Another factor is the negatively skewed distribution of Matrix test scores, which would tend to exaggerate decreases among the less intelligent groups. But as 1 point at a score of 42 corresponds approximately to $1\frac{1}{2}$ points at 32, this, too, could only account for a small proportion of the differential decline.

Table XI brings out the negative correlation between intelligence and decline in yet another way. It shows the percentages of each total age group obtaining very high Matrix scores. In spite of the steadily decreasing mean, the proportion scoring

TABLE XI. PERCENTAGE DISTRIBUTIONS IN DIFFERENT AGE GROUPS

Score	Years of Age						
	16-17	18-19	20-24	25-29	30-34	35-39	40+
55-60 ..	0.33	0.36	0.32	0.35	0.15	0.04	0.00
50-54 ..	3.72	3.71	4.02	2.55	1.69	1.29	0.69
35-49 ..	65.85	62.64	59.30	52.99	48.50	42.80	33.13
5-34 ..	30.10	33.29	36.36	44.11	49.66	55.87	66.18
Mean ..	37.68	37.13	36.43	34.66	33.74	32.40	30.30

55 or over remains steady (or even rises) till 29 years, and only then starts to drop rapidly. Again the proportion scoring 50-54 remains steady (or rises) to 24 years, and then drops. Even here, however, we cannot be certain that the very high scorers were not volunteers from reserved occupations for such high-grade jobs as pilots, though there is no direct evidence for this explanation.

V. DISCUSSION

Our data tend to support though they do not allow us to prove, the hypothesis of differential decline. If confirmed it is of the greatest psychological and educational importance, since it suggests that men who make most use of their intelligence retain it best, and that men in unskilled and labouring occupations tend to lose the mental capacities they possessed when they left school. Norris (1940) has shown clearly that educational attainments which are not used in daily life are rapidly forgotten after leaving school, and the performance of adults on verbal intelligence tests has been

found by Lorge to be higher the more recent and more extensive their schooling, even when 14-year-old intelligence level is held constant (cf. Garrett, 1946). Now the Matrix test is one of the purest tests of g available, hence it should not be affected by education or by the exercise in manipulating verbal and numerical symbols which, say, Clerks and Electricians get. But, as the present writer has shown elsewhere (Vernon, 1938), a test of this type is just as much affected by what he has called 'test sophistication' as is a verbal test. Scores of superior adults are raised by about $\cdot 6\sigma$, corresponding to five points on the Matrix test, by familiarity with test construction and practice in taking tests. It may then be this factor of habituation to pencil and paper work as much as g which declines among those whose schooling is long past, and whose jobs involve little or no verbal and numerical exercise. To draw a sharp distinction, however, between the capacity to do tests and g or intelligence is hardly sound. For g can only be recognised in so far as it expresses itself through some medium, and until a medium is discovered which is unaffected by sophistication and education, we are entitled to claim that g itself is subject to their influence.

It appears then that effective intelligence, or g , does not remain constant from the age of 15 or so throughout early adulthood, as is widely assumed, but rises, or more often falls, in different groups, depending upon the education or other forms of training, and the intellectual activities which these groups indulge in or neglect. We see also that adult test norms which purport to apply to men or women in general are misleading, since they would no longer be valid if any change occurred in the age, occupational, or educational make-up of the populations tested, or in the degree of familiarity of the testees with the tests. Finally we need to investigate more fully the problems of transfer of training among adults. There seems to be more truth than most psychologists have admitted in the lay view that "brains" get "rusty" from lack of "practice." Just what types of education, and occupational and avocational experience, most effectively stimulate g is a matter of the utmost concern to democracy.

VI. SUMMARY

1. The scores of 89,764 candidates for the Royal Navy on the Progressive Matrices Test (20 mins. version) were tabulated under twelve occupational categories, also according to age and area of Great Britain. It is shown that the Wren Recruiting Assistants who undertook the occupational classification carried it out with adequate consistency.

2. Since these three factors overlap, and the numbers in the various sub-groups were very uneven, special measures had to be taken to establish their separate contributions. Analyses were carried out, and the means for the several groups were obtained, according to two assumptions :

- (i) that all three factors operate independently ;
- (ii) that Occupation and Area are inter-connected, but that the age distribution is constant in all occupational-area sub-groups.

3. Variance due to all factors is highly significant, though the difference between the most intelligent occupational groups (Clerks and Electrical Workers) and the least intelligent (Labourers and Farm Workers) is smaller than is found with verbal tests. Area differences are very small when the occupational distribution is held constant, partly because of the mixed constitution of each area. An estimate is obtained of the separate and of the joint contribution of these two factors.

4. The decline with age from 17 to over 40 years is very marked and, in this sample, is almost linear ; but the comparability of successive age groups is uncertain. The rapidity of decline differs significantly in different occupations, and appears to be related to intelligence level, being greatest in the least intelligent occupations. More-

over, in spite of the steady drop in mean score with age, the proportion of very high scorers (50 points and over) remains constant or even rises from 16 to 25 or 30 years, suggesting that the most able retain their intelligence best. The psychological implications of this finding are pointed out.

REFERENCES

1. Cattell, R. B. (1943): 'The measurement of adult intelligence.' *Psychol. Bull.*, XL., 153-193.
2. Garrett, H. E. (1946): 'The effects of schooling upon IQ: A note on Lorge's article.' *Psychol. Bull.*, XLIII, 72-76.
3. Miles, C. C., and Miles, W. R. (1932): 'The correlation of intelligence scores and chronological age from early to late maturity.' *Amer. Journ. Psychol.*, XLIV, 44-78.
4. Norris, K. E. (1940): *The Three R's and the Adult Worker*. Montreal, McGill University Press.
5. Raven, J. C. (1939): 'The R.E.C.I. series of perceptual tests: An experimental survey.' *Brit. Journ. Med. Psychol.*, XVIII, 16-34.
6. Vernon, P. E. (1938): 'Intelligence test sophistication.' *Brit. Journ. Educ. Psychol.*, VIII, 237-244.

APPENDIX

MATRIX TEST DISTRIBUTIONS IN OCCUPATIONAL, AGE, AND AREA GROUPS

The age distribution is held constant in occupational and area groups. Numbers are reduced to a total of 10,000.

	Occupational Groups											
	1 Clerks	2 Drivers	3 Builders	4 Pre- cision	5 Sheet Metal	6 Oper- ators	7 Wood- work	8 Electric	9 Trade	10 Farm Work	11 Labour	12 Mates
55+	20.8	0.0	0.0	3.1	0.9	0.4	0.2	2.3	0.9	0.0	0.4	0.2
50+	162.7	3.7	10.1	36.8	12.4	10.7	14.0	30.6	20.8	2.8	9.5	8.9
45+	355.4	40.6	69.2	177.8	77.6	71.5	85.2	90.9	112.8	15.5	72.6	49.3
40+	382.7	120.0	172.3	337.5	173.3	152.9	216.4	150.4	261.1	42.4	227.5	125.4
35+	212.2	156.7	208.2	316.5	200.2	162.0	200.3	112.5	263.4	57.9	328.0	139.3
30+	93.9	130.8	143.2	187.2	129.2	119.0	132.2	75.4	183.1	53.8	303.3	110.3
25+	41.1	100.7	102.2	108.6	73.5	76.6	61.8	28.9	105.7	41.7	238.1	79.5
20+	15.1	53.2	51.2	45.2	32.0	39.6	28.2	16.3	59.9	30.4	181.5	43.0
15+	6.1	24.3	20.9	13.1	10.0	16.6	7.0	3.4	26.6	14.3	82.2	29.6
10+	2.9	9.9	10.3	6.4	3.9	11.4	1.7	2.3	8.6	8.6	48.3	10.9
5+	0.1	1.1	1.4	0.8	0.0	0.3	0.0	0.0	1.1	0.6	8.6	1.6
N	1,293	641	789	1,233	713	661	747	513	1,044	268	1,500	598

	Age Groups						
	16-17	18-19	20-24	25-29	30-34	35-39	40+
55+	11.9	12.3	2.2	1.6	0.9	0.3	0.0
50+	134.3	125.7	27.7	11.5	10.0	10.9	2.9
45+	512.8	459.5	90.3	45.8	44.1	50.7	15.2
40+	959.0	840.0	169.8	97.7	112.0	134.8	48.6
35+	907.2	823.0	148.6	95.5	131.4	176.1	75.4
30+	553.4	549.5	100.5	73.3	119.1	173.0	92.6
25+	314.5	319.9	71.5	58.9	85.3	137.6	70.7
20+	143.8	166.9	46.0	32.0	47.6	96.9	62.4
15+	50.5	61.1	22.9	22.1	26.7	36.7	34.1
10+	23.1	29.4	9.1	9.9	14.7	23.6	15.4
5+	1.5	2.7	0.4	2.7	1.2	4.4	2.7
N	3,612	3,390	689	451	593	845	420

	Area Groups									Total
	Birm- ingham	Bristol	Derby	Glasgow	Liver- pool	London	Man- chester	New- castle	South- ampton	
55+	2.0	2.0	1.6	1.5	1.9	11.4	5.1	0.7	3.0	29.2
50+	28.0	21.3	19.3	27.7	23.8	108.8	55.1	14.9	24.1	323.0
45+	108.2	90.4	78.6	124.2	95.6	355.3	194.6	71.2	100.3	1,218.4
40+	228.9	192.2	148.8	268.5	193.4	640.6	351.4	135.7	202.4	2,361.9
35+	252.3	206.1	145.4	286.9	199.8	584.3	333.0	141.2	208.2	2,357.2
30+	190.4	147.4	95.5	213.3	153.3	381.5	234.4	99.4	146.2	1,661.4
25+	125.8	106.1	61.9	149.3	97.1	231.3	131.6	60.8	94.5	1,058.4
20+	63.6	72.8	36.9	98.5	54.7	114.2	65.7	40.3	48.9	595.6
15+	29.0	33.3	15.9	46.7	23.0	47.3	21.8	16.3	20.8	254.1
10+	16.1	17.3	6.1	28.2	11.3	17.0	13.9	6.8	8.5	125.2
5+	2.7	2.1	1.0	4.2	1.1	1.3	1.4	0.7	1.1	15.6
N	1,047	891	611	1,249	855	2,493	1,408	588	858	10,000

THE ASSOCIATION BETWEEN AGE AND SCORE IN THE PROGRESSIVE MATRICES TEST

By PATRICK SLATER

Government Social Survey

I. *Reasons for the Enquiry.* II. *Data Collected.* III. *Results.* IV. *Summary.*

I. REASONS FOR THE ENQUIRY

The enquiry reported here was made for the Directorate of Selection of Personnel, War Office, during 1943 ; and a report on it was circulated within the Directorate in January, 1944, but not published. At that time all men entering the Army were given a standard set of tests. The results, together with other information obtained by interview and questionnaire, were used for deciding the training to be given to each man. Training recommendations were made by Personnel Selection Officers in accordance with instructions issued by the Directorate, which included specifications of minimum test standards for different duties. Men failing to reach the standard were not to be recommended for the duty.

The method most generally used for deciding what standards to set was to follow up sample groups of men undergoing training for the duties concerned ; to compare the records of their success in training with the observations made on them during the selection procedure ; to calculate multiple regression equations between the data and the criterion of success ; and to fix standards in terms of the data which were thus shown to carry the greatest weight. Similar methods have, of course, been widely used for similar problems. An exact parallel to the methods employed at the Directorate can be found, if desired, in reference (1). The assumptions they involve are general, and their usefulness in particular instances can be measured. Thus even when they are not exactly true they may be retained until others measurably more efficient can be discovered.

The chronological age of the men was recorded and used in this way. Its correlation with other variables, such as test scores or assessments of success, was calculated by the usual product-moment formula. When the correlation thus obtained was used in computing multiple regression equations, the weight assigned to age rested on the assumption that the regression of the other variables on age is rectilinear. Carried to its logical conclusions, this assumption is absurd. It would indicate infants as the most desirable recruits for occupations in which age is negatively associated with success, and centenarians for occupations in which it is positively associated. However, it may approximate sufficiently closely to the truth during the period of life when most men are in military service to be preferred to any more elaborate one.

It is generally believed that intelligence increases proportionately with chronological age up to an age somewhere between fifteen and twenty ; that after maturity it continues at a stable level until the age of about thirty ; and that after that a decline becomes perceptible which continues slight until about fifty, but becomes thereafter more marked. The relation cannot be described more accurately so long as it is discussed in terms of a characteristic so nebulously defined as intelligence. Thus it seems difficult to determine how far the normal one-one relation found between mental and chronological age in childhood is from being merely an artefact of the scale in which mental age is measured. In the description of their revision of the Binet test Terman and Merrill (2) reveal the extent of their uncertainty thus : " The

number of three-year-olds who test at four is approximately equal to the number of six-year-olds who test at eight or nine-year-olds who test at twelve. This suggests (though it does not prove) that the mental growth between three and four years is of about the same magnitude as that which occurs between six and eight or between nine and twelve."

We are on much safer ground, if we abandon the attempt to describe the relation between age and intelligence and describe instead the relation between chronological age and score in a specific test. Terman and Merrill give precise information for ages up to eighteen. The most precise information about older people is that given by Miles and Miles (3). They used a specially printed form of the Otis Self-Administering test of intelligence, given with a fifteen-minute time limit, on a sample of 823 subjects aged from seven to over ninety. A useful general discussion is provided by Freeman (4).

Naturally even less is known about the development, maturity, and decline of other psychological traits concerning the definition of which agreement is less general—traits, for example, which might be supposed to underlie the observations collected during the selection procedure and the subsequent assessments of success recorded by the Directorate. There was no need to consider any such problems in order to assess how much error was involved in treating the regression of the observed variables on age as rectilinear; indeed, any such attempt would have led into devious arguments involving superfluous uncertainties, and would thus have provided unnecessarily unreliable conclusions.

Of all the variables concerned, the Progressive Matrices test, which was regularly given with a time limit of twenty minutes, consistently showed the closest degrees of association with age. If no excessive errors could be found to be involved in treating this association as rectilinear, it seemed fair to argue that no serious error would be involved in treating the associations of the other variables with age similarly. This is the problem examined here.

II. DATA

Records of 100 people born in each year from 1901 to 1925 inclusive were drawn from data collected in the Army during the second half of 1943. To ensure comparability between the years, the restriction was imposed that each sample of 100 should have the same distribution by Educational Standard. Table I shows the composition of the samples in this respect.

TABLE I

Educational Standard						Number of cases
Code	Definition					
11	Degree or equivalent qualification	}	}	}		2
12						
21						
22						
31	Schools Certificate or equivalent	}	}	}		3
32						
41	Full-time education until 16 but no certificate	}	}	}		4
42						
51	Any full or part-time education after 14	}	}	}		10
52						
61	Primary education only, normal attainments					8
71	Partly illiterate					53
71	Partly illiterate					14
81	Illiterate					1
Total						100

The definitions given are for the first of the figures in each code number, which indicate levels of educational attainment. Of the figures in the second place, 1 indicates education with a literary or academic bias, 2 indicates a technical bias. Elementary or lower educational attainments were not considered sufficient to warrant a suffix of 2 in any case. Partial illiteracy (E.S. 71) was the standard assigned to anyone, whatever his education, who was in one of the bottom two grades (lowest 30 per cent. approximately) on both the Arithmetic and Verbal tests. The distribution given in this table corresponds, as closely as could be estimated, with the distribution prevailing in the serving Army at that time.

III. RESULTS

Table II shows the average score for each sample.

TABLE II

Sample of 100 men aged :	Average score on Progressive Matrices :
18	39.89
19	40.17
20	37.02
21	35.94
22	35.15
23	37.24
24	37.69
25	36.57
26	33.95
27	36.27
28	37.57
29	36.55
30	35.93
31	36.46
32	33.61
33	34.55
34	33.86
35	36.48
36	36.77
37	35.02
38	33.45
39	33.15
40	33.56
41	32.08
42	30.94
Average all ages	35.5948

And Table III gives particulars for comparing the variation between the groups with the variation within them.

TABLE III.

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Square
Variance between age-groups	11,788.723	24	491.197
Variance within groups (error variance)	191,135.810	2475	77.227
Total variance of test scores	202,924.533	2499	—

The variance between the groups defines the limit to which test scores can be treated as dependent on age : no hypothesis can account for more than 5.81 per cent. of their total variance. Although the association between age and score is thus not very great during this period of life, it is shown to be significant by the ratio of the two mean squares ($F=6.36$) ; the null hypothesis has a probability of less than .01.

Polynomial regression equations to the fifth degree were fitted to the series of observed averages given in Table II, thus testing the improvement in goodness of fit obtainable by taking successively more complicated hypotheses concerning the relationship between age and score, starting from the simple one that it is rectilinear. The results obtained are analysed in Table IV.

TABLE IV. IMPROVEMENTS IN FIT OBTAINED BY SUCCESSIVE INCREASES IN THE ORDER OF THE POLYNOMIAL FITTED

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Squares
Amount accounted for by First order equation (rectilinear regression)	7,143.581	1	7,143.581**
Increase accounted for by Second order equation	47.069	1	47.069
Third	1,264.025	1	1,264.025**
Fourth	41.396	1	41.396
Fifth order	275.572	1	275.572
Residual variance between groups	3,017.080	19	158.794**
Total variance between groups	11,788.723	24	—

The mean squares marked ** give values of F (when compared with the error variance shown in Table III) with probabilities less than .01 ; other probabilities exceed .05 and can be treated as insignificant.

Details can be easily checked from the figures given. The main conclusions are :

1. 60.6 per cent. of the total variance between groups can be accounted for by rectilinear regression. This amount can be increased significantly by fitting a polynomial of the third order ; an additional 11.1 per cent. can be accounted for thus, i.e., 71.7 per cent. altogether.

These figures may also be compared with the total variance of the scores (Table III). Variations between the age groups account for 5.81 per cent. of the total. When the regression is treated as linear, 3.52 per cent. is accounted for ; the error involved in treating the regression as linear is 2.29 per cent. of the total. If a regression equation of the third order is used, 4.17 per cent. is accounted for, and the error involved is reduced to 1.64 per cent. The gain is only 0.65 per cent.

2. Further significant increases cannot be obtained by fitting polynomials of the fourth or fifth order.

3. The residual variance between groups remains significant, whether a straight regression line is fitted to the data, or a polynomial of any of the first five orders. None of the fits is really good.

In terms of correlations, the standardized coefficient of rectilinear regression, .1876, is, by definition, equivalent to a correlation coefficient (r) ; the ratio of the variance between groups to the total variance is equal to the square of the correlation ratio (η), which is therefore .2410.

Since the association between age and the other variables under consideration had already been found to be generally less than with scores on this test, and since data obtained from follow-up studies were less likely to be suited for exact analysis than those used in this specially designed experiment, it was concluded that for such purposes the regression of age on all other variables might safely continue to be treated as if it were linear.

The Appendix gives expressions from which interpolations can be made in equations of the first five orders.

This report is a result of collective work. I am deeply grateful to the War Office for permission to publish.

IV. SUMMARY

1. The enquiry reported here was made for the Directorate of Selection of Personnel, War Office, during 1943, to discover whether any serious errors were involved in treating the association between age and psychological tests, etc., as rectilinear during the period of life when men were serving with the Army. As the Progressive Matrices Test, which was given with a twenty-minute time limit, was found to be more closely associated with age than other tests, scores on it were made the object of a special study. A sample of 2,500 men was taken, consisting of 100 men of each year of age from 18 to 42 inclusive. Each group was matched in educational attainment, and representative of the standards found in the serving Army at that time.

2. Their scores were found to decline gradually with age from an average of just about 40 to just over 30. No hypothesis can account for more than 5·81 per cent. of the total variance of the scores, for this is the total amount attributable to differences of age; the remaining 94·19 per cent. is due to differences between men of the same age.

3. Rectilinear regression accounts for 3·52 per cent.; a polynomial equation of the third order would account for an additional 0·65 per cent. As the sample is a large one, this increase can be shown to be significant, but it was considered to be too small to warrant abandoning the simple assumption of rectilinearity for general purposes.

REFERENCES

1. Halstead, H., and Slater, Patrick: 'An experiment in the vocational adjustment of neurotic patients.' *Journ. of Mental Science*, Vol. XCII, No. 388 (July, 1946).
2. Terman, L. M., and Merrill, M. A.: *Measuring Intelligence*. G. G. Harrap and Co., 1937.
3. Miles, C. C., and Miles, W. R.: 'The correlation of intelligence scores and chronological age, from early to late maturity.' *Amer. Journ. of Psych.*, Vol. 44 (1932).
4. Freeman, F. S.: *Individual Differences*. G. G. Harrap and Co., 1936.

APPENDIX

The general form of the equation to be fitted is :

$$\left(\frac{x-\bar{x}}{\sigma_x}\right)\beta_x + \left(\frac{x^2-\bar{x}^2}{\sigma_{x^2}}\right)\beta_{x^2} + \left(\frac{x^3-\bar{x}^3}{\sigma_{x^3}}\right)\beta_{x^3} + \dots \text{etc.} \dots = \frac{m-\bar{m}}{\sigma_m}$$

The necessary values for interpolating in this equation can be found as follows :

$$\bar{x} = \text{mean age} = 30$$

To find the average score, m , of a sample of men aged x , take $x-\bar{x} = d$, and the following values :

$$\begin{aligned}\sigma_x &= 7.211103 \\ x^2 - \bar{x}^2 &= d^2 - 52 \\ \sigma_{x^2} &= 46.398276 \\ x^3 - \bar{x}^3 &= d^3 \\ \sigma_{x^3} &= 734.081739 \\ x^4 - \bar{x}^4 &= d^4 - 2152.8 \\ \sigma_{x^4} &= 6432.514793 \\ x^5 - \bar{x}^5 &= d^5 \\ \sigma_{x^5} &= 90672.655525 + \\ \sigma_m &= 9.009429 \\ \bar{m} &= 35.5948\end{aligned}$$

and the following values of β :

For an
equation
of the

		β_x	β_{x^2}	β_{x^3}	β_{x^4}	β_{x^5}
1st	order	.187625	.0	.0	.0	.0
2nd	..	.187625	-.015230	.0	.0	.0
3rd	..	.005565	-.015230	.198431	.0	.0
4th	..	.005565	-.063747	.198431	.050573	.0
5th	..	.140393	-.063747	-.217073	.050573	.298966

BOOK REVIEWS

Multiple Factor Analysis. By L. L. Thurstone. University of Chicago Press. Pp. xx + 535. \$7.50.

The subtitle describes this book as "a development and expansion of *The Vectors of Mind*." Most of what was essential in the former volume is retained; but an additional purpose of the new and amplified edition is "to make available the results of work that has been done during the past decade in the Psychometric Laboratory at the University of Chicago." By this is to be understood work bearing on the methods of factor-analysis—on its theory and techniques: with the concrete results and the psychological conclusions reached by the aid of those methods the present volume does not profess to deal. Nearly all Thurstone's recent contributions to this aspect of the subject, most of them hitherto accessible only in isolated articles or papers, are now brought together in a systematic and connected exposition. Accordingly, whereas the first edition comprised only ten chapters, the present volume now extends to more than twice that number. On the other hand, the appendix outlining the working procedure for the centroid method is omitted: instead we are promised a "work book on multiple factor analysis" in the near future.

Thurstone believes that "the brevity of *The Vectors of Mind* was probably responsible for much of the controversy about multiple factor-analysis." And in particular he finds that "the fundamentally simple notion which I called 'simple structure' has been twisted around by some critics into the most inconceivable nonsense." One of his chief aims, therefore, is to clarify his arguments by fuller exposition and illustration, and above all to explain in greater detail what he means by 'simple structure,' and how it is to be attained.

The definition of 'simple structure' is virtually the same as before: "If a set of r hyperplanes of dimensionality $(r-1)$ exists such that each trait vector is in one or more of the hyperplanes, then the combined configuration of the trait vectors and the reference vectors will be called a simple structure." The criteria for a *unique* simple structure, however, are now raised from three to five. The object of the new requirements is to increase still further the number of zero entries; but at the same time they make it still more probable that the simple structure thus determined will be oblique rather than orthogonal.

The difference between this type of factor matrix and that favoured by most factorists in this country can be easily stated. Burt, for example, 30 years ago suggested that the results of applying cognitive and scholastic tests to school children could commonly be expressed by a matrix of uncorrelated factors, of which the first (accounting for most of the variance) was a positive general factor, and the remainder a series of 'group factors' (slightly overlapping), such as a verbal and a literary factor, a numerical factor, a manual or practical factor, etc. Thurstone prefers to analyse such data into terms of group factors only. The group factors he then obtains are much the same as when a general factor is included; but they are correlated. Thus the general factor no longer manifests itself explicitly as such; it operates covertly by producing a correlation between the factors forming the simple structure.

This situation leads Thurstone to a full discussion of 'second-order factors'—a concept that did not appear in the earlier edition. Second-order factors are defined as "factors that are obtained from the correlations of the first-order factors." One resolution that he offers yields " $(r+1)$ linearly dependent factors" consisting of second-order general and group factors: and he adds that "this type of resolution is preferred by some students, who use the reference frame G, a, b, c, d and e " (second-order factors) "because it is orthogonal rather than the frame G, A, B, C, D , and E which is oblique." He concludes his chapter on the subject by suggesting that "an interesting application of second-order factors is an attempt to reconcile three theories of intelligence, namely, Spearman's theory of a general intellectual factor, Godfrey Thomson's sampling theory, with what he calls sub-pools, and our own theory of correlated multiple factors."

British readers will turn with interest to Prof. Thurstone's discussion of factorial studies carried out in this country before he himself entered the field. In his preface he states that "when multiple-factor analysis was introduced in 1931, . . . papers were mainly concerned with Spearman's central intellectual function g ; group factors were regarded as

the disturbers of the general factor." And later in the volume he again states that "before multiple factor analysis was introduced . . . these other factors . . . were described as 'disturbers' of the tetrad equation, and avoided by eliminating similar tests from the battery." And he continues: "this type of factorial analysis is still current among British psychologists; most of them have not yet adopted the multiple factor methods."

These statements need (to say the least) to be supplemented. The first set of equations ever published which proposed to reduce observable correlated variables to terms of hypothetical uncorrelated components (the equations suggested by Pearson in 1901) were essentially equations for an analysis in terms of multiple factors. Indeed, it was precisely the reduction of this analysis by multiple common factors to an analysis in terms of a single common factor only, which Spearman subsequently proposed, that formed the chief target for attack by Karl Pearson and his followers in the early days of factorial work. A glance at Brown and Thomson's *Essentials of Mental Measurement* should suffice to demonstrate that by no means all British psychologists accepted Spearman's "type of factorial analysis." And, as may be seen from the appendix to the Board of Education's 1924 *Report on Psychological Tests of Educable Capacities* (not to mention numerous publications in the *British Journal of Educational Psychology* and elsewhere), few teachers and educationists in this country were willing to agree that group factors were mere "disturbers" and as such to be "eliminated." However, if Thurstone's main point is that British psychologists regard group factors as "secondary in importance" to the general cognitive factor (i.e., as contributing less to the total variance), then his account of the difference between the two "schools" is so far justified.

Other problems which are dealt with more fully than in the earlier edition are those of assessing communalities and of making rotations. A new and important chapter deals with "the effects of selection"; and here full acknowledgment is made of the work of Godfrey Thomson. There is, too, a most suggestive chapter on "configurations and factor patterns." The entire book is characterized by the logical clarity and lucidity of exposition which distinguish all Thurstone's writings. The new diagrammatic and photographic illustrations, and the simple concrete examples worked out in the course of the discussion, make it possible for the student, who has read the mathematical introductions, to follow the arguments with ease. The volume as a whole makes a most valuable addition to the literature of factor-analysis; and must be both read and digested by every psychologist who attempts to make anything more than a routine contribution of his own to the subject. C. B.

Probit Analysis. A statistical treatment of the sigmoid response curve, by D. J. Finney. Camb. Univ. Press, 1947. Pp. xiii + 256. 18s.

"There is a need to-day" says Dr. Finney in his preface, "for books in which the specialized statistical methods appropriate to certain branches of science will be discussed in sufficient detail to enable biologists to appreciate them and apply them to their own problems." His book fulfils this task admirably and explains and illustrates probit analysis by description, by examples, by supplying the necessary tables, and by giving, in an appendix, its mathematical bases. In addition, Dr. Finney gives an account, which is of peculiar interest to psychologists and especially to the present reviewer, of the history of the method. In its modern form it was devised and developed mainly by J. H. Gaddum (1933) and C. I. Bliss (1934), but much earlier it had under another name, the *Constant Process*, and with another nomenclature, been developed by psychophysicists. It is a remarkable instance of the watertight compartments of specialized science; for those of us who earlier had a hand in developing the statistics of the *Constant Process* remained quite ignorant of the fact that Probit Analysis was practically the same thing. The accident that Dr. Finney was a friend and correspondent of Dr. D. N. Lawley, who at the time was on my staff, led to our enlightenment.

Probit Analysis in its modern form has mainly developed in connection with research on insecticides, and similar researches where the application of some chemical, or other treatment, results in the destruction of some insects or germs or what not. A weak application kills only a few; stronger applications kill a percentage increasing with the strength of the chemical or whatever it is. A diagram of the percentage mortality against the strength

of the poison gives the "sigmoid response curve" of Dr. Finney's subtitle, a curve rising at first slowly, then more steeply, and finally flattening out as it approaches 100% asymptotically. It is exactly the same kind of curve as the psychophysicist obtained when in "threshold" experiments he set off the number of positive responses against the value of the stimulus. Except for modern refinements (mainly the use of Maximum Likelihood instead of Least Squares), the psychophysicist's subsequent statistical treatment was the same as in probit analysis. That treatment consists in fitting to the sigmoid curve a theoretical curve which is the integral of a normal distribution. A "probit" (splendid selling name !) is simply a normal deviate with 5 added to it to avoid negative numbers. I fancy the word is made up from the first and last syllables of *probability unit*. In the *Constant Process* the percentage of positive answers, and the value of the stimulus, were inserted in an equation containing the normal integral, and there were as many such equations as there were stimuli used in the experiment. The percentage of positive answers corresponds to the percentage *kill* in an insecticide experiment, the size of the stimulus to the strength of the insecticide. The equations have to be solved for the central point (the 50% *kill*) and the standard deviation. There are too many equations and normal equations have to be formed, which is not possible unless the equations are linear. These equations are therefore turned into linear ones—and this is the essential point of both the *Constant Process* and *Probit Analysis*—by equating the upper limits of the integrals to the normal deviate (the "probit" in the modern language) corresponding to the percentage. Finney rightly gives the credit for inventing the method to Fechner (1860), and describes the recognition by Müller (1879) of the need for certain weighting of the equations, and by Urban (1909, 1910) for further modifications thereof. The combined Müller-Urban weights were given by Urban in the *Psychological Review* in 1910 and are to be found in Appendix I of *The Essentials of Mental Measurement*, first edition (Wm. Brown) 1911, and in all subsequent editions (Brown and Thomson). That table exactly corresponds to Table 4 on Finney's page 32, except for a constant factor .637 and for the fact that Urban tabulates the weights against the percentage *kill*, while the probit form tabulates them against the probit or normal deviate corresponding to that *kill*. Thus at a 69% *kill* Urban gives $w = .914$. Finney's Table 4 gives .581 (i.e. $.914 \times .637$) at probit 5.5, that is at a normal deviate of 0.5, which in a Gaussian curve corresponds to a *kill* of 69%. The constant factor .637 is of course immaterial in weights, and is due to the simplification in the Müller-Urban table of making the largest weight unity.

It has already been mentioned that the Probit Method uses the Method of Maximum Likelihood where the psychophysicists used the Method of Least Squares (the former having not then been invented). The other big difference is that the Probit Method goes carefully into the statistical significance of the estimates arrived at, giving their confidence limits. This was a refinement beyond Fechner and Müller. But Urban made an attempt to do something of the sort. All credit is due to him for seeing the necessity and for trying, but in fact he erred grievously in his confidence estimates. It fell to the present reviewer to correct this (in the *Brit. Journ. Psychol.*, May 1914) by a sound but very cumbersome and inelegant form of calculation, to which Finney gives full acknowledgment, saying "His revision put the constant process into a form very similar to that of the probit method to-day, except that the weights were taken from . . . the observations and not from . . . a provisional line, and that his formulæ for the standard errors of the parameters were much more complicated."

It is clear that students of statistical psychology must possess and study Finney's book. He has himself applied Probit Analysis to the results of mental tests (*Psychometrika*, March 1944) following a paper by George Ferguson (*Ibid*, Feb. 1942) of the University of Toronto but earlier on the Moray House staff, using the *Constant Process*. Ferguson wrote this in ignorance of the probit work. Still more remarkable, D. N. Lawley, also unaware thereof, proceeded to apply the method of maximum likelihood to Ferguson's problem (*Proc. Roy. Soc. Edin.* March 1943), that is to advance the *Constant Process* along the path already taken by the insecticide people. As I said at the outset, if Finney and Lawley had not been correspondents this independent breaking of the same path might have gone on still longer.

I have used my space mainly on a comparison interesting to psychologists and helpful to them. I will only say further about Finney's book that it is most welcome, most "efficient," and deserving our "confidence."

Godfrey Thomson

INTERNAL AND EXTERNAL FACTOR ANALYSIS

By M. S. BARTLETT

Department of Mathematics, University of Manchester

- I. *The Relation of Factor Analysis to External Forms of Canonical Analysis.*
- II. *Discussion of Some Tests of Significance.*

I. THE RELATION OF FACTOR ANALYSIS TO EXTERNAL FORMS OF CANONICAL ANALYSIS

Introductory Remarks. The papers by Burt (1) and Godfrey Thomson (2) in the first issue of this *Journal* refer respectively to factor analysis and to Hotelling's "most predictable criterion." It may be useful to examine further these two types of statistical analysis in a brief survey of what might in a generalized sense be termed factor analysis. The most predictable criterion to which Godfrey Thomson refers, in the correlation or regression relations between two batteries of tests, is the first of a set of canonical variables or factors into which one battery is transformed, the second battery having a corresponding set of canonical variables. This analysis of the mutual relations between two sets of variables (Hotelling (3)) I have discussed and illustrated at some length elsewhere (4); the first point I wish to stress here is that the canonical transformation of one group of variables into "factors" is made with reference to an external criterion, viz., the other group of variables, and for this reason I shall refer to it as an external factor analysis, in contrast with the internal canonical or factor analysis of a single set of variables.

Factor Analysis of Mental and Physical Measurements. The difference may perhaps be easily grasped when the variables fall naturally into two groups. If a set of mental tests and also a set of physical measurements are made on the same group of persons, the internal analysis of the mental scores will yield the ordinary factor analysis, as described by Burt (1). A corresponding internal analysis of the physical measurements will yield a factor analysis of physical structure (see, for example, the analysis recently described by Burt and Banks (5)). The mutual relations (if any) of the two groups of measurements would be examined by means of an external factor analysis. I believe attempts at establishing such relations between physical and mental measurements have met with little success, though some of these have concentrated on mental testing of "intelligence," whereas any mental factors correlated with physique might be expected to lie (for example, in Kretschmer's theory of types) more in temperament.* In any case it should be noticed that if significant factors emerged from such an analysis, they would not necessarily be identical with, or have the same order of importance, as the internal factors; the "shape factor" found by Burt and Banks, if it remained, might well take first place over the factor of "size" in the set of physical measurements.

* Small correlations between physique and temperament have been established by Burt (see, for example, (11)), to whom I am indebted for further information on these investigations. He remarks that while some Italian writers seem to have assumed body-type would be correlated with intelligence, most English psychologists would agree in expecting correlations (if any) to be with temperament.

The distinction made here between internal and external factor analysis enables me to clarify a remark I made some years ago on the statistical analysis of physical and mental measurements and which I may perhaps quote (see end of footnote, (6) p. 98) :

“ The chief value of the idea of psychological types must lie, not in an arbitrary classification of men into two or more grades, but in a demonstration that independent classifications of this kind (by, for example, the estimation of factors) are correlated. If from physical tests it was found possible to classify men according to a factor corresponding to Kretschmer's contrast of pyknic and asthenic types (the main general factor emerging from such tests would presumably be one merely of pure size), and it was also found possible to make an independent classification of temperament by means of mental tests, a significant correlation between the physical and mental factors would have provided a statistical justification in ordinary persons of Kretschmer's theory of types. Such a procedure would, it will be noticed, have conformed to orthodox test and factor techniques ; it might in fact be regarded as a particular case of the general factor method of showing that two tests, or two series of tests, are not independent, but are better explained in terms of an underlying “ factor,” though the case is a very special one because of the natural dichotomy of physical and mental measurements.”

The publication, shortly after this remark was first written, of Hotelling's method of analysis between two groups of variables enables the “ natural dichotomy of physical and mental measurements ” to be taken account of in the form of factor analysis employed.

An External Form of Tests—Persons Analysis. Having emphasized this difference between internal and external analysis (a difference which includes the obvious logical point usually, though not invariably, perceived by mental testers that an internal analysis on any set of tests does not in itself imply any relations whatsoever with an external criterion like subsequent performance), my second aim is to show that the difference may sometimes be a matter of approach or of interpretation, and that for some purposes a set of test scores for a group of persons might validly be analysed by an external form of analysis. Burt (1) has drawn an analogy of factor analysis with analysis of variance ; the analogy must not be pushed too far, however, for the former treats the different tests as different variables, while the latter would attempt to treat them all in terms of one variable. It is in fact instructive to consider also the interpretation which proceeds in the reverse direction, where we treat not only the tests but the persons as different variables. The suggestion has of course previously been made (Stephenson (7)) of making a factor analysis of persons, but also has this limitation of having to regard different tests all in terms of one variable ; statistically, the interpretation becomes more orthodox if we consider tests as one set of variables and persons as a second. It is shown below that an external analysis may be made between tests and persons, provided that we have an independent assessment of variability “ within persons,” such as is afforded by repetition of any tests on the same persons.

The use of “ reliability coefficients ” as weights is well known in psychology, but this does not mean that the analysis being considered here corresponds to any standard usage. It has more points of similarity with another form of external analysis that has been suggested by Godfrey Thomson (8), who, by correlating tests with a duplicate set, investigated what combination of tests or “ factor ” was most reliable. The present analysis will have the property of isolating the factors which discriminate most between persons. Its practical value I regard as limited, because there is no guarantee whether the best discrimination will correspond to general or to specific ability ; though it might be added that as the repetition of the tests separates in time, the consistency of factors tends to be replaced by the “ permanency ” of factors, and this has perhaps greater interest.

My primary intention in this paper, however, is to contrast with these external forms the internal form of factor analysis, with its rather arbitrary weighting following

from the standardization of the test scores, and to indicate the difference in the appropriate tests of significance in the two cases.

Suppose we have obtained scores for r persons in p tests, with t "repetitions" for each test. For example, the tests might be marks from ordinary school examinations repeated (with different questions) at the end of more than one term; in this case, any average differences in scores due to lapses of time would naturally be isolated in the subsequent analysis, and for definiteness the form of the analysis is indicated for this case. It may be noted that analysis of variance technique can be regarded as an important simple case of multiple regression technique with degenerate values for the independent variables. In regression terminology we may correlate tests with persons, the variable representing any person having the score 1 for that person and 0 for all others; but this analysis will be identical with a multi-variate analysis of variance (analysis of variance and covariance, cf. (4)) of tests "between" and "within" persons. This analysis of variance and covariance would have the following structure* by rows:

Between persons	..	$r-1$	degrees of freedom
Within persons	..	$(t-1)(r-1)$	degrees of freedom
Between times	..	$t-1$	degrees of freedom
Total	..	$tr-1$	degrees of freedom

The first two items, which are the two we are interested in, have a total of $t(r-1)$ degrees of freedom, of which $r-1$ correspond to differences between persons (the degrees of freedom for persons are only $r-1$ instead of r as the practice of measuring test scores from their means automatically eliminates the scores of the "average person," representing one degree of freedom). If we denote the set of sums of squares and products in the analysis corresponding to the first and second rows by the (symmetric) matrices L and M , the canonical analysis of tests is associated with the roots of the determinantal equation:

$$|L - \lambda(L + M)| = 0. \quad (1)$$

This equation is an example of the general determinantal equation associated with the external analysis of one set of variables with another set. To recapitulate the form of the equations in general (for the method of deriving them, see, for example, 4 §3), suppose that after eliminating the general mean or other irrelevant degrees of freedom the deviations of the 'dependent' set of variables are represented by the matrix S_2 (one row representing each variable). Any canonical variable or 'factor' of S_2 determined by S_2 's relation with the corresponding matrix S_1 for the other set of variables is determined from the equations:

$$(C_{21}C_{11}^{-1}C_{12} - \lambda C_{22}) \mathbf{a} = 0. \quad (2)$$

where λ is a root of

$$|C_{21}C_{11}^{-1}C_{12} - \lambda C_{22}| = 0, \quad (3)$$

$C_{22} = S_2 S_2'$, $C_{11} = S_1 S_1'$ being the matrices of internal sums of squares and products for the variables of S_2 (S_2' denoting the transpose of S_2) and S_1 respectively, and $C_{21} = S_2 S_1'$ the matrix of cross sums of products. The vector quantity $\mathbf{a}$ in equation (2) denotes the column of coefficients which when multiplied to the variables of S_2 defines the factor. The corresponding factor from S_1 is determined by the column of coefficients $C_{11}^{-1}C_{12}\mathbf{a}$. The matrix $C_{21}C_{11}^{-1}C_{12}$ in equations (2) and (3) represents the

* The structure is similar to that of the analysis given in Table I of (4).

sums of squares and products corresponding to the part of S_2 which is predictable in terms of S_1 , and with this interpretation the relation of equation (1) to the more general equation (3) becomes clearer.

Equation (3) and thus equation (1) have the same theoretical structure as Godfrey Thomson's equation (5) in his paper (2); the values of the roots are not altered by the reduction of the entries to correlation coefficients, which, however, would be more artificial in the present problem. In Godfrey Thomson's example, the second non-zero root would be the other root in his quadratic equation for λ . The number of significant factors, and the relative importance of the different variables, may be investigated by the general methods outlined in my paper (4); (cf. also II below).

Corresponding to each significant root in equation (1), there will be a linear combination of tests or a "factor." Further, to each such factor there will correspond a linear combination of the variables representing persons, typifying a fictitious person correlating most highly with such a particular test factor. It will be perceived that the analysis represented by equation (1) is exact in the sense that all significance is based on comparison with the consistency or reliability of the tests.

Relation with Godfrey Thomson's External Factor Analysis for Reliability. Before we examine equation (1) further to see in what circumstances it would be equivalent to an internal factor analysis, my reference to Godfrey Thomson's analysis of reliability (8) makes it perhaps desirable to demonstrate the theoretical difference between his and the above analysis. In terms of the problem discussed by Godfrey Thomson in his exposition (2) of the correlation of two different batteries of tests, he considers in his earlier paper the special case where the second battery is a "duplicate" of the first. This means that the internal correlation matrix of the second battery is (theoretically, sampling fluctuations not being considered) the same as the first, $R = \{r_{ij}\}$, say; and also that the cross-correlation matrix R_0 between the two batteries is identical with R except for the appearance in the diagonal of the correlations r_i between duplicates of the same test T_i . Godfrey Thomson's general equation (5), paper (2), becomes then in this special case

$$| R_0 R^{-1} R_0 - \lambda R | = 0. \quad (4)$$

For comparison, in equation (1) we suppose t , the number of repetitions of the same test, to be 2, and also allow the number of persons to become large so that sampling fluctuations vanish. We have for the covariance between any two tests T_i and T_j , even when we first pool all the test scores available for each of these tests, the value $r_{ij}\sigma_i\sigma_j$, where σ_i is the standard deviation of test T_i . The covariance between the quantities $(T_i + T'_i)/\sqrt{2}$ and $(T_j + T'_j)/\sqrt{2}$, which correspond to the "between persons" degrees of freedom, where T_i and T'_i refer to the two repetitions of the same test, is $4 r_{ij}\sigma_i\sigma_j/(\sqrt{2})^2 = 2 r_{ij}\sigma_i\sigma_j$. This excludes the case $i=j$, the variance of $(T_i + T'_i)/\sqrt{2}$ being $(2+2r_i)\sigma_i^2/(\sqrt{2})^2 = (1+r_i)\sigma_i^2$. Hence, dividing equation (1) by the total number $2(r-1)$ degrees of freedom before letting r become large, and dividing the i -th row and column by σ_i in order to reduce to correlations, we obtain

$$| \frac{1}{2}(R + R_0) - \lambda R | = 0 \quad (5)$$

in contrast with equation (4). If we write $R - R_0 = D$, equation (5) may be written equivalently

$$| \frac{1}{2}D - \lambda' R | = 0 \quad (5a)$$

where $\lambda' = 1 - \lambda$. In both equations (4) and (5) the occurrence of any test T_1 , say, for which $r_1 = 1$ automatically gives a root $\lambda = 1$; further, if D is small, equation (4) may be written approximately

$$| R - 2D - \lambda R | = 0, \quad (6)$$

or, if $\lambda' = 1 - \lambda$,

$$| 2D - \lambda' R | = 0. \quad (6a)$$

In certain situations the roots and hence the analyses corresponding to equations (4) and (5) are thus closely related, but in general the equations and roots will be different.

Relation of the Tests-Persons Analysis with Internal Factor Analysis. Returning to equation (1), we note that the actual values of the roots depend on the numbers of degrees of freedom for L and M , $r-1$ and $m=(t-1)(r-1)$ respectively. It is convenient for the following discussion to change to the new scale μ , where

$$\mu = m\lambda/(1-\lambda) \quad (7)$$

so that μ satisfies the equation

$$|L - \mu M/m| = 0 \quad (8)$$

or

$$|L - \mu V_e| = 0 \quad (9)$$

where V_e is an estimated variance and covariance matrix of "error." If m becomes large, or if the true error matrix V is assumed known on other grounds, equation (9) becomes

$$|L - \mu V| = 0. \quad (10)$$

If in this equation we assume further that "errors" in different tests are uncorrelated (an assumption not made in the sample equation (1)), the unknown value μ appears only in the diagonal terms multiplied by these error variances, and the form of equation familiar in internal factor analysis begins to appear. Writing $(q-1)\mu = \nu$, and reducing L to the corresponding correlation matrix R , equation (10) may be written

$$|R - \nu v_{ii}| = 0 \quad (11)$$

where v_{ii} is the ratio of a diagonal element of V in (10) to the corresponding variance obtained from L .

From the standpoint of equation (11) the internal analysis obtained from the equation

$$|R - \nu| = 0 \quad (12)$$

is equivalent to the two assumptions that (i) the ratios of the diagonal elements of V to the true variances corresponding to L may be assumed equal, and that (ii) any sampling fluctuations of the actual diagonal elements of L may be neglected.* In practice equation (12) is usually adopted (in psychology) in preference to (11), even when they differ; for example, equation (11) implies an infinite weight for a test which is completely reliable, even if it represents a purely specific factor (it was noted that in this it is similar to Godfrey Thomson's weighting of a battery of tests for reliability). This merely means that (11) and (12) are isolating factors from the tests for different purposes. The existence of the two equations does, however, stress the difficulty when in (12) the criterion based on assessment of error is abandoned—namely, that any other weighting must be to some extent arbitrary. Of course, even in equation (12) the degree of psychological "reality" of any factors emerging from the analysis must be a matter of opinion conditioned by further evidence, and is not established from the data alone. Two or more tests correlate highly if they are saturated with a common factor; but the question whether this is because the tests are all minor variants of the same test, or because the common factor is indeed general to all tests, the scores themselves can never answer.

* I do not of course wish to imply by this remark that equation (12) has no other justification; internal factors are statistically definable as the transformed variables in terms of which a set of variables can most economically be expressed; and equation (12) corresponds to an internal analysis into principal components (see Hotelling (10)).

II. DISCUSSION OF SOME TESTS OF SIGNIFICANCE

The χ^2 Approximation. A good approximate criterion that may be used for the significance of the smallest k roots λ_i ($i=1 \dots k$) in the analysis represented by equations (2) and (3) is to calculate

$$\chi^2 = - \left\{ n - \frac{1}{2}(p+q+1) \right\} \sum_{i=1}^k \log_e(1-\lambda_i) \quad (13)$$

where n is the total number of degrees of freedom ($t(r-1)$ in the analysis of variance problem discussed) and where p and q ($p \leq q$) are the numbers of 'dependent' and 'independent' variables (e.g., p tests, and $q=r-1$ degrees of freedom among persons). It has been stressed (cf. 4) that the corresponding allocation of $k(q-p+k)$ degrees of freedom to χ^2 in (13) will only provide a valid approximate test when the larger roots $\lambda_{k+1} \dots \lambda_p$ correspond to well-determined factors (pairs of orthogonal vectors in the sample space of the two sets of variables, corresponding to each of these larger roots, then tend to coalesce, and the variation of the k and $q-p+k$ remaining variables in the two sets can be considered in the reduced sample space of $n-p+k$ degrees of freedom).

When n (or m of equation (7)) becomes very large, the λ_i tend to zero, but the quantities $n\lambda_i$, or μ_i in (7) remain finite. Equation (13) thus becomes, in the limit,

$$\chi^2 = \sum_{i=1}^k \mu_i, \quad (14)$$

an appropriate criterion for an equation like (10).

The Significance of the Ratio of the Two Smallest Roots. In both (13) and its limiting form (14) the existence of an "error" component in our analysis has enabled us to have an absolute criterion of the significance of the roots λ_i , and this will in general be available for testing significance in an external canonical analysis. If for any reason it is not, and roots are judged by their *relative* magnitude (e.g., if the error variances in (10) were unknown but assumed equal), then the above allocation of degrees of freedom should *not* be made for testing the significance of the smallest root but one against the smallest (assuming the latter insignificant), for we cannot logically consider the second smallest root simultaneously as well determined and yet be testing its significance. However, from the joint distribution* of the two smallest roots on the hypothesis that these are, in contrast with the others, arising merely by chance, the significance level for the ratio of the larger to the smaller can be approximately calculated. Critical values were calculated in this way and are given in Table I. They are based on the formula :

$$P\{x\} = \left(\frac{2\sqrt{x}}{1+x} \right)^s \quad (15)$$

for the significance level $P\{x\}$ of the ratio x . This formula may be shown to be exact when there are only two roots altogether, when $s=q-1$; and will be approximately correct under the above more general conditions, when $s=q-p+1$.

It is further stressed that the above formula does *not* apply to equation (12), which must be considered independently. Here also the allocation of degrees of freedom for the purpose of testing significance by ordinary analysis of variance tests

* For the distribution of the roots, see, e.g., Wilks (9).

is invalid. As noted by Hotelling (10, p. 434), the ratio in the case $p=2$ is in fact expressible in terms of the correlation coefficient r_{12} between the two tests, and the significance of x follows at once from the relation

$$x = \frac{1 + |r_{12}|}{1 - |r_{12}|} \quad (16)$$

(Hotelling's distribution for x on p. 434 is wrong by a factor of 2 owing to his not ensuring that $x \geq 1$, but this error does not extend to the numerical example on p. 435, since the significant value of r_{12} that he adopts is in effect that of the absolute value $|r_{12}|$.) Again, the degrees of freedom for x in equation (16) should be reduced from $s=q-1$ ($=r-2$) to $q-p+1$ when other larger roots have been eliminated. Significant values of x corresponding to equations (12) and (16) are included for comparison in Table I.

TABLE I. SIGNIFICANT VALUES OF x (RATIO OF TWO SMALLEST ROOTS)

Degrees of Freedom s	External Analysis *		Internal Analysis †	
	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$
1	1,598	39,998	648	16,259
2	77.99	397.75	39.00	199.00
3	27.44	84.17	15.43	47.46
4	15.83	37.97	9.60	23.15
5	11.17	23.19	7.15	14.94
6	8.74	16.52	5.82	11.07
7	7.28	12.83	5.00	8.89
8	6.30	10.55	4.43	7.50
9	5.61	9.02	4.03	6.54
10	5.09	7.92	3.72	5.85
12	4.36	6.46	3.28	4.91
14	3.88	5.54	2.98	4.30
16	3.53	4.91	2.76	3.87
18	3.28	4.45	2.60	3.56
20	3.07	4.10	2.46	3.32
25	2.72	3.50	2.23	2.90
30	2.48	3.12	2.07	2.63
35	2.30	2.85	1.96	2.44
40	2.19	2.66	1.88	2.30
45	2.09	2.51	1.81	2.19
50	2.01	2.39	1.75	2.10
60	1.89	2.21	1.67	1.96
70	1.80	2.10	1.60	1.86
80	1.74	1.98	1.55	1.79
90	1.68	1.91	1.52	1.73
100	1.64	1.84	1.48	1.68

* Corresponding to equations (10) and (15). † Corresponding to equations (12) and (16).

It will be perceived that compared with the external analysis a smaller value of the ratio x is significant, and not a larger value, as would have been inferred if standardization were assumed to reduce the effective number of degrees of freedom. This is probably due, at least in part, to the standardization of scores in the internal analysis having the effect of restricting the variability of x . If we do compare these values with ordinary analysis of variance tests (remembering that the appropriate variance ratio has degrees of freedom s in the denominator, and $s+2$ in the numerator, and must be multiplied by $(s+2)/s$ to convert it to a ratio x of sums of squares) we

find that there is no particular correspondence with the values of x in Table I for an external analysis, though the agreement is fair for an internal analysis (without any further reduction of degrees of freedom for standardization in the latter case). For example, we should have from the standard significance levels of the variance ratio F , for $s=5, 10$, and 20 corresponding values of x :

s	5	10	20	40	80
$x(P=0.05) \dots$	6.83	3.49	2.31	1.76	1.44
$x(P=0.01) \dots$	14.63	5.65	3.19	2.21	1.73

The use of these values would for an internal analysis not be far wrong for $s=5-10$, though they would give rather too few significant values as s becomes large.

Numerical Illustrations. To illustrate these various tests, suppose that Godfrey Thomson's fictitious correlation table on p. 30 of his paper (2) had been calculated from 20 persons. Then (one degree of freedom eliminated with the mean) we have in formula (13) :

$$n=19, \quad p(\leq q)=2, \quad q=3.$$

We find also, in addition to his $\lambda_1=0.6850$, that $\lambda_2=0.4530$. Hence the significance of the relation of the two batteries will be indicated by

$$\chi^2 = -16 \{ \log_e(1-\lambda_1) + \log_e(1-\lambda_2) \}$$

with degrees of freedom $(q-p+3)+(q-p+1)=4+2=6=pq$; that is, by the χ^2 table :

	D.F.	χ^2
λ_1	4	$-16 \log_e(1-0.685)=18.48$
λ_2	2	$-16 \log_e(1-0.453)=9.65$
Total	6	$-16 \log_e[(1-0.685)(1-0.453)]=28.13$

We conclude from the significance of the second component that the factors corresponding to the first root do not exhaust the significant relations between the two batteries.

To illustrate the use of equation (14) (in general a cruder approximation than (13)) suppose that an analysis of the relation of ten physical measurements and six tests of temperament on 1,001 persons ($n=1,000$) had given the six ($p \leq q=6, q=10$) roots $\lambda_1=0.2, \lambda_2=0.1, \lambda_3=0.08, \lambda_4=0.025, \lambda_5=0.02, \lambda_6=0.004$. Remembering that $n\lambda_i$ is roughly a χ^2 with $q+p+1-2i$ degrees of freedom, we see that the largest three roots are obviously significant, while for the smallest three we have the approximate analysis :

	D.F.	χ^2
λ_4	9	$1,000 \times 0.025 = 25.0$
λ_5	7	$1,000 \times 0.02 = 20.0$
λ_6	5	$1,000 \times 0.004 = 4.0$
Total	21	$1,000 \times (0.025 + 0.02 + 0.004) = 49.0$

We should conclude that λ_4 and λ_5 , though small, are also significant, while λ_6 is insignificant. If, however, the value of n had not been recorded and, assuming λ_6 a random deviation from zero, we calculate the ratio λ_5/λ_6 , this is only 5, whereas for degrees of freedom $s=5$, we should require from Table I a ratio of 11.17 for $P=0.05$ significance. This stresses the insensitiveness of the ratio test, especially when the number of degrees of freedom s is small, and it should obviously not be employed in the usual situation arising for an external analysis when the significance of the λ_i can be directly ascertained.

For an internal analysis, on the other hand, the significance of the roots is necessarily equivalent to a test of their ratios. Thus in Burt's Table XV (see (1)), the value of x for the last two factors is $0.1593/0.0058=27.5$, whereas for $s=2$ ($q-p+1=(r-1)-p+1=5-4+1$) the $P=0.05$ significance level calculated from the relation (16), or equivalently read off from Table I, is 39.0.

REFERENCES

1. Burt, C. : 'A comparison of factor analysis and analysis of variance.' *Brit. Journ. Psychol. (Statistical Section)*, I, 1947, 3.
2. Thomson, Godfrey H. : 'The maximum correlation of two weighted batteries.' *Brit. Journ. Psychol. (Statistical Section)*, I, 1947, 27.
3. Hotelling, H. : 'Relations between two sets of variables.' *Biometrika*, XXVIII, 1936, 321.
4. Bartlett, M. S. : 'Multivariate analysis.' *Journ. Roy. Stat. Soc. (Suppl.)*, IX, 1947, 176.
5. Burt, C., and Banks, Charlotte : 'A factor analysis of body measurements.' *Annals of Eugenics*, XIII, 1947, 238.
6. Bartlett, M. S. : 'The statistical conception of mental factors.' *Brit. Journ. Psychol.*, XXVIII, 1937, 97.
7. Stephenson, W. : 'The inverted factor technique.' *Brit. Journ. Psychol.*, XXVI, 1936, 344.
8. Thomson, Godfrey H. : 'Weighting for battery reliability and prediction.' *Brit. Journ. Psychol.*, XXX, 1940, 35.
9. Wilks, S. : *Mathematical Statistics* (Princeton, 1943).
10. Hotelling, H. : 'Analysis of a complex of statistical variables into principal components.' *Journ. Educ. Psychol.* XXIV, 1933, 417 and 498.
11. Burt, C. : 'The analysis of temperament.' *Brit. Journ. Med. Psychol.*, XVII, 1938, 158.

NOTE ON THE RELATIONS OF TWO WEIGHTED BATTERIES

By GODFREY THOMSON

Moray House, Edinburgh

I. *Errata.* II. *Mutual External Factorization.*

I. ERRATA

I regret that there is an arithmetical slip in my previous paper, which luckily does not affect the subsequent calculations. On p. 31 the column and the row have been divided by 12.4 instead of by its square root, and a similar blunder occurs on p. 33. But as I immediately expressed the weights as *relative* weights (which are correct) the blunder escaped my notice. On p. 31 the marginal correlations (instead of 0.105, etc.) should have been

0.369 0.767 0.397

and the absolute values of the weights (instead of 0.107, etc.) should have been

0.376 0.786 —0.143

I should have employed a check by seeing that the inner product of these two sets (which is now 0.685) was equal to the largest root of the equation for λ .

Similarly at the bottom of p. 33 the marginal correlations should be

0.514 0.800

and the absolute values of the weights

0.231 0.708

whose inner product, as a check, is again 0.685, the root λ_1 . In passing I may note that the exact values of the two roots to six places are $\lambda_1 = 0.684955$ and $\lambda_2 = 0.453030$.

I repeat that fortunately the relative weights, which alone I used, are correct.

II. MUTUAL EXTERNAL FACTORIZATION

I take this opportunity of completing the calculations using the other roots (0.453 and zero), in illustration of Professor M. S. Bartlett's paper preceding this note. In effect, each of the two batteries dictates a factorial analysis of the other battery into orthogonal factors, the first pair of factors being as near to each other as the two separate spaces will allow; or in other words, the first pair of factors being the two batteries each so weighted and summed as to give maximum correlation, $\sqrt{\lambda_1} = 0.828$, with the other. In this part of the calculation, in my previous paper, the largest root of either

$$| CA^{-1}C' - \lambda B | = 0$$

$$\text{or } | C'B^{-1}C - \lambda A | = 0$$

was used.

If we use the next root of either equation, $\lambda_2 = 0.453$, and exactly the same procedure as before, we obtain weights

0.565 —0.331 0.116 for the larger battery,

and 0.710 —0.458 for the smaller.

With these weights two axes are defined which are orthogonal to the previous two, and as near together as this condition will permit. Their correlation is $\sqrt{\lambda_2}=0.673$.

The third root λ_3 of the cubic is zero, and using this root in exactly the same way we arrive at weights, for the larger battery, in the ratio of any row of the adjugate of $C'B^{-1}C$ itself, such as

$$0.080724 \quad 0.028644 \quad -0.130200$$

(though I cannot see that any absolute values, as distinct from proportional values, can be identified in this case). These weights define an axis in the three-space of the larger battery which is orthogonal to the two previous axes in that space, and to the whole of the two-space of the smaller battery. The correlations of the 3+2 axes or factors with one another are in fact

	α_1	α_2	α_3	β_1	β_2
α_1	1	.	.	.828	.
α_2	.	1	.	.	.673
α_3	.	.	1	.	.
β_1	.828	.	.	1	.
β_2	.	.673	.	.	1

PREDICTION OF A COMPLEX CRITERION AND BATTERY RELIABILITY

By E. A. PEEL

University of Durham

I. *Introduction.* II. *Battery Correlation.* III. *Maximum Battery Reliability :* (a) *Thomson's Method ;* (b) *Suggested New Method ;* (c) *Equivalence of the Methods.* IV. *Worked Examples :* (a) *Maximum Prediction of an Arbitrarily Weighted Complex Criterion ;* (b) *Maximum Battery Reliability.* V. *Summary.*

I. INTRODUCTION

A battery of tests may be used to predict an external criterion by weighting each test of the battery in the familiar way (1, Chap. VI), so that the error of estimate of the criterion is a minimum. This fundamental condition is symbolized in the statement that

$$\sum_i (x_i - \hat{x}_i)^2$$

shall be a minimum, where x_i are the criterion scores and $\hat{x}_i$ are the criterion scores as estimated by the weighted battery of tests used for prediction. Thus aptitude for an engineering trade might be predicted by a battery composed of tests of manual dexterity, three dimensional space form perception, non-verbal and verbal intelligence. The criterion would consist of some engineering performance which formed a valid assessment of the ability and could be scored unambiguously. These scores would be symbolized by x_i and the weighted sum of test scores would yield the estimated aptitude scores $\hat{x}_i$.

Many of the criteria, however, which we meet in problems of educational and vocational guidance are not sufficiently well defined by a single measure and more than one assessment might need to be made in order to obtain a more complete account of the ability. If we take the example of ability in engineering trades we may consider that such ability is revealed in the technical school course by success in metalwork, technical drawing, and woodwork. The criterion is now composed not of a single assessment but of a battery of three assessments, and the problem outlined in the first paragraph becomes one of choosing test weights for the battery of tests to give maximum prediction of a *complex* criterion composed of a *battery* of assessments.¹ Hotelling (2 and 3) first considered and provided a solution for this type of problem, although he was primarily concerned with weighting the criterion, that is, with finding what kind of criterion the tests would best predict. In this paper the writer is in addition concerned with the converse problem of so weighting a battery of tests that they best predict a certain complex criterion. The weights assigned to the components of the complex criterion may be selected quite arbitrarily² on educational, industrial,

¹ In order to avoid ambiguity, assessments are regarded throughout this paper as the components of the complex criterion, although tests can form part of a criterion. Tests refer to the material used for *predicting* such a criterion.

² Arbitrarily in the sense of being determined on grounds other than the mathematical principle of least squares.

or psychological grounds. Thus we might decide that in the above battery of assessments, metalwork, technical drawing, and woodwork should be weighted respectively in the ratios 3 : 2 : 1 for engineering trades associated with, say, the construction of locomotives. We can then calculate test weights to give *maximum* prediction of the complex criterion as defined by the arbitrarily weighted battery of assessments. This prediction is the largest possible under the conditions imposed by the arbitrary selection of assessment weights. It would not be, however, maximum prediction in the sense defined by Hotelling who *finds* sets of weights for both tests and assessments which yield mathematically the highest possible correlation between the batteries of tests and assessments. As Thomson¹ has observed, however, the criterion so weighted might not be the one in which we are interested, for we may wish to assign arbitrary weights to the assessments on other grounds.

As prediction is measured by battery correlation it is proposed to consider the correlation between two batteries in greater detail. We can envisage three 'degrees' of correlation between the batteries of tests and assessments. In the first place we may calculate correlation between an arbitrarily weighted test battery and an arbitrarily weighted assessment battery. This correlation is uniquely defined and no maximal problem exists. It would not be generally the largest correlation possible between the batteries. Secondly, we may arbitrarily assign weights to the assessment battery and calculate the test weights which give maximum battery correlation with the complex criterion defined by the arbitrarily weighted assessments. Thirdly, we may calculate the weights for both batteries which give the mathematically largest correlation between the two batteries. These weights are determined purely on mathematical grounds and not from any psychological consideration. A useful special case of this last maximum battery correlation is that of maximum reliability. Thomson (4) first enquired into this problem and provided a solution for obtaining the weights which give maximum battery reliability. In such a case we are concerned not with batteries of tests and assessments but with correlating a battery of tests with the second application of the same battery.

At the end of his paper Hotelling (2) remarks that it is possible that the most predictable criterion will have little correlation with many of the original criteria and Thomson (4) states the problem more explicitly when he observes that the weights which give maximum prediction are not generally the same as those which give maximum reliability. He then propounds the problem of finding a set of weights which will "make battery reliability equal to prediction and both as large as possible under this condition" (4, p. 365), and provides an example involving a battery of two tests and a single criterion, from which he obtains the required test weights.

It is the object of this paper to reconsider Hotelling's method for maximum prediction and Thomson's application to battery reliability with a view to developing a method for obtaining tests weights which give maximum prediction of a given complex criterion and to deriving a shorter method for calculating maximum battery reliability (5). The notation of matrix algebra is used and is probably familiar to those who have read Thomson, Burt, and others in this country.

Worked examples are also provided illustrating the two methods set out by the writer. The possibility of providing a general solution of the problem of finding test weights which make reliability and prediction equal and both as large as possible is still under investigation and is not considered in this paper.

¹ In correspondence with the writer. This correspondence has been very helpful in clarifying the writer's views on the two kinds of maximum prediction.

II. BATTERY CORRELATION

We first derive the three 'degrees' of correlation between two batteries. If we consider a set of a variates and a set of b variates, the matrix of correlation coefficients between the $a+b$ variates may be symbolized as

$$\begin{bmatrix} R_{aa} & R_{ab} \\ R_{ba} & R_{bb} \end{bmatrix}$$

Furthermore if we assign the vector of weights u to the a variates and the vector w to the b variates, we obtain the following pooling square :

	u'	w'
u	R_{aa}	R_{ab}
w	R_{ba}	R_{bb}

On application of the product moment correlation coefficient the battery correlation between the batteries of a and b variates is given by r , where

$$r = \frac{u'R_{ab}w}{u'R_{aa}u \times w'R_{bb}w} \quad (1)$$

When equation (1) is applied to tests and criterion assessments, a corresponds to the assessments and u to their arbitrary weights, b refers to the tests and w to their weights. The weights u and w are selected quite freely without any intention of maximizing r .

Suppose now we fix the assessment weights u according to some psychological consideration and we enquire what values of the test weights w would give the best prediction of the complex criterion formed by the arbitrarily weighted assessment battery. To answer this question we require the maximum battery correlation which is obtained as follows.

We select the weights w , so that the complex criterion formed by the weighted sum of a variates, $u'a$, has least mean square error with its estimate formed by the weighted sum of the tests, $w'b$. This means that

$$\Sigma(u'a - w'b)^2$$

has to be a minimum. The sum may be written as the matrix product

$$(u'A' - w'B')(Au - Bw)$$

where A and B are the respective matrices of assessment and test scores, corresponding to the a and b variates. On differentiating with respect to the column vector w and equating the partial derivatives to zero we obtain the equation

$$u'A'B - w'B'B = 0$$

and since $A'B = R_{ab}$ and $B'B = R_{bb}$

$$w' = u'R_{ab} R_{bb}^{-1} \quad (2)$$

The values of the weights w given by equation (2) are those which should be assigned to the tests b in order to obtain maximum prediction of the arbitrarily weighted criterion $u'a$.

The value of the maximum prediction is obtained by substituting for w in equation (1) by the expression (2). When this is done the maximum prediction is given by

$$r = \sqrt{\left(\frac{u'R_{ab} R_{bb}^{-1} R_{ba} u}{u'R_{aa} u} \right)} \quad (3)$$

an expression which contains only the observed matrices of correlation coefficients R_{aa} , R_{bb} , and R_{ab} and the arbitrarily assigned values of the assessment weights u .

The battery correlation given by equation (3) represents the second 'degree' of correlation. It has importance psychologically in connexion with weighted batteries of tests used for educational or vocational guidance. A worked example will follow later.

Although r of equation (3) is the maximum correlation when the vector of weights u has been arbitrarily chosen, it can be increased to a final maximum by selecting u in such a way that the expression $u'R_{ab} R_{bb}^{-1} R_{ba} u$ is maximized under the condition that $u'R_{aa} u$ remains constant, say, at unity. This leads us to the third 'degree' of correlation between the two batteries, that is, to maximum prediction in the sense defined by Hotelling (2). In order to effect the maximization we have to select undefined multipliers so that

$$u'R_{ab} R_{bb}^{-1} R_{ba} u - \lambda u'R_{aa} u$$

is a maximum, that is, on forming the partial derivatives with respect to u' , so that

$$R_{ab} R_{bb}^{-1} R_{ba} u - \lambda R_{aa} u = 0 \quad (4)$$

The condition for the non-trivial solution of these equations is that

$$| R_{ab} R_{bb}^{-1} R_{ba} - \lambda R_{aa} | = 0 \quad (5)$$

We solve this equation for its largest root λ_1 , in order to find the values of u which make r a maximum. It can be shown that $\lambda_1 = r_m^2$, where r_m is the maximum prediction. The assessment weights u which give this maximum prediction are in the ratio of the elements of any row of the matrix

$$\text{adj} (R_{ab} R_{bb}^{-1} R_{ba} - \lambda_1 R_{aa})$$

The test weights w follow by substituting this value of u in equation (2).

The value r_m is that which is obtained by making the 'best' of the correlation coefficients represented by the matrix $\begin{bmatrix} R_{aa} & R_{ab} \\ R_{ba} & R_{bb} \end{bmatrix}$.

III. MAXIMUM BATTERY RELIABILITY

(a) *Thomson's Method.* We have symbolized the tests by the set of b variates and to preserve the logical unity of the argument we shall continue to use this symbol for the first application of the tests and symbolize the second application by the β variates. For the matrix of correlations we shall therefore have

$$\begin{bmatrix} R_{bb} & R_{b\beta} \\ R_{\beta b} & R_{\beta\beta} \end{bmatrix} \text{ giving a pooling square } \begin{array}{c|cc} & u' & w' \\ \hline u & R_{bb} & R_{b\beta} \\ w & R_{\beta b} & R_{\beta\beta} \end{array}$$

We can continue to use u and w for the weights without any ambiguity.

Under the special condition set out by Thomson (4) all the submatrices are square and $R_{bb} = R_{\beta\beta}$ and $R_{b\beta} = R_{\beta b}$. Furthermore

$$R_{b\beta} = R_{bb} - I + \Lambda \quad (6)$$

where Λ is a diagonal matrix, having individual test reliabilities as diagonal entries.

This last fact is irrelevant to the theoretical development of Thomson's method and the method devised by the writer, but it is of use in the computation of maximum battery reliability.

Thomson considers the reliability

$$r = \sqrt{\left(\frac{u' R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b} u}{u' R_{bb} u} \right)} \quad (7)$$

and maximizes $u' R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b} u$ under the condition $u' R_{bb} u = 1$. His method is strictly analogous to that used by Hotelling leading ultimately to solving the equation

$$| R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b} - \lambda R_{bb} | = 0 \quad (8)$$

for its largest root λ_1 in order to obtain the weights u which give maximum reliability r_m . The root λ_1 is equal to r_m^2 and the weights u are in the ratio of the elements of any row of the matrix

$$\text{adj} (R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b} - \lambda_1 R_{bb})$$

If b , the first application of the tests, were maximized then we should be led to the equation

$$| R_{\beta b} R_{bb}^{-1} R_{bb} - \lambda R_{\beta\beta} | = 0 \quad (9)$$

for finding the weights w . Since $R_{bb} = R_{\beta\beta}$ and $R_{\beta b} = R_{b\beta}$ equations (8) and (9) lead to the same result and u is equal to w , which is what we should expect when considering reliabilities.

(b) *Suggested New Method.* In order to calculate maximum prediction and maximum reliability by equations (5) and (8) we must first compute the triple matrix products $R_{ab} R_{bb}^{-1} R_{ba}$ and $R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b}$. Although the former cannot be avoided Thomson (4, p. 361) has simplified the latter a little by applying equation (6). The product $R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b}$, however, can be avoided altogether by using the following method.

We have $u = w$, $R_{bb} = R_{\beta\beta}$ and $R_{b\beta} = R_{\beta b}$ and on making the necessary substitutions in the equation

$$\rho = \frac{u' R_{b\beta} w}{\sqrt{u' R_{bb} u \times w' R_{\beta\beta} w}}$$

which gives a measure of reliability ρ and is analogous to equation (1), we obtain

$$\rho = \frac{u' R_{b\beta} u}{u' R_{bb} u} \quad (10)$$

It is evident from equation (10) that maximum reliability can be obtained by maximizing $u' R_{b\beta} u$ under the condition $u' R_{bb} u = 1$, leading to a solution of the equation

$$| R_{b\beta} - \lambda R_{bb} | = 0 \quad (11)$$

for its largest root λ_1 which in this case is equal to the maximum reliability ρ_m . The weights u which are associated with maximum reliability are given by any row of the matrix

$$\text{adj} (R_{b\beta} - \lambda_1 R_{bb})$$

The advantages of this method are obvious when we consider that the triple product $R_{b\beta} R_{\beta\beta}^{-1} R_{\beta b}$ of equations (8) and (9) has been replaced by the single matrix $R_{b\beta}$ in equation (11). The determinant on the left hand side of (11) has also certain advantages over the corresponding determinant in (8), for its sign and value when $\lambda = 1$ can be calculated easily. This proves of use when finding the largest root λ_1 .

(c) *Equivalence of the Methods.* r_m and ρ_m are the same if $r=\rho$ (equations (6) and (10)), since the condition for maximization, namely, $u' R_{bb} u = 1$, is the same in each case. On squaring the expressions (6) and (10) we have to show that

$$(u' R_{b\beta} u)^2 = (u' R_{bb} u) (u' R_{b\beta} R_{bb}^{-1} R_{\beta b} u) \quad (12)$$

or, after expanding that,

$$u' (R_{b\beta} u u' R_{\beta b}) u = u' (R_{bb} u u' R_{b\beta} R_{bb}^{-1} R_{\beta b}) u \quad (13)$$

In order to prove that the scalar terms on each side of (13) are equal it is necessary and sufficient to show that $R_{bb} u u' R_{b\beta} R_{bb}^{-1} R_{\beta b}$ is the transpose of $R_{b\beta} u u' R_{bb}$. On post-multiplying each of these terms by the product $R_{b\beta}^{-1} R_{bb}$ and recalling that $R_{b\beta}$ is symmetrical and non-singular, we obtain the products $R_{b\beta} u u' R_{bb}$ and $R_{bb} u u' R_{b\beta}$. Since R_{bb} , $u u'$, and $R_{b\beta}$ are all symmetrical it follows that

$$R_{bb} u u' R_{b\beta} = (R_{b\beta} u u' R_{bb})' \quad (14)$$

whence equation (13) is proved and therefore $r=\rho$ and hence

$$r_m = \rho_m.$$

IV. WORKED EXAMPLES

(a) *Maximum Prediction of an Arbitrarily Weighted Complex Criterion.* We have to calculate the vector of test weights w which are given by equation (2)

$$w' = u' R_{ab} R_{bb}^{-1}$$

and the maximum prediction by equation (3)

$$r = \sqrt{\left(\frac{u' R_{ab} R_{bb}^{-1} R_{ba} u}{u' R_{aa} u} \right)}$$

Let us take an example in which the criterion is composed of two assessments and the test battery is composed of three tests. The data are collected from a research recently carried out by the writer on predicting technical ability. The assessments consist of woodwork and technical drawing and the tests of two assembly tests and one non-verbal paper test. The battery of correlation coefficients is as follows :

			u_1	u_2	w_1	w_2	w_3
u u' R_{aa} w' R_{ab}		u_1	1	·638	·314	·379	·301
		u_2	·638	1	·229	·258	·294
	=	w_1	·314	·229	1	·518	·390
w R_{ba} R_{bb}		w_2	·379	·258	·518	1	·372
		w_3	·301	·294	·390	·372	1

Prediction of a Complex Criterion

We first calculate $R_{ab} R_{bb}^{-1}$ using Aitken's method (see Thomson (1) p. 361).

					Sum
1	·518	·390	—1		·908
·518	1	·372		—1	·890
·390	·372	1			·762
·314	·379	·301		—1	·994
·229	·258	·294			·781
·7317	·1700		·5180	—1	·4197
1·0000	·2323		·7079	—1·3667	·5735
·1700	·8479		·3900		·4079
·2163	·1785		·3140		·7088
·1723	·2047		·2290		·5731
·8084			·2697	·2323	—1
					·3104
1·0000			·3336	·2874	—1·2370
·1283			·1609	·2956	
·1723			·1303	·1905	
					·3840
					·5848
					·4931
$R_{ab} R_{bb}^{-1}$			·1181	·2587	·1587
			·0728	·1410	·2131
					·5355
					·4269

and we have for the test weights

$$\begin{aligned} w_1 &= \cdot1181 u_1 + \cdot0728 u_2 \\ w_2 &= \cdot2587 u_1 + \cdot1410 u_2 \\ w_3 &= \cdot1587 u_1 + \cdot2131 u_2 \end{aligned} \quad (15)$$

Suppose now we decide to weight the assessments, woodwork, and technical drawing, arbitrarily in the ratio 1 : 1, then by (15)

$$w_1 = \cdot1909, w_2 = \cdot3997, w_3 = \cdot3718$$

and the maximum prediction [equation (3)] is obtained by first calculating

$$\begin{aligned} u' R_{ab} R_{bb}^{-1} R_{ba} u &= [\cdot1909 \quad \cdot3997 \quad \cdot3718] \begin{bmatrix} \cdot314 & \cdot229 \\ \cdot379 & \cdot258 \\ \cdot301 & \cdot294 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \cdot5795 \end{aligned}$$

and then

$$u' R_{aa} u = [1 \quad 1] \begin{bmatrix} 1 & \cdot638 \\ \cdot638 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 3\cdot276$$

and finally

$$r = \frac{\sqrt{\cdot5795}}{\sqrt{3\cdot276}} = \cdot4206$$

If we decide to weight the assessments in the ratio 1 : ·6, then by equations (15)

$$w_1 = \cdot1618, w_2 = \cdot3433, w_3 = \cdot2866,$$

and

$$\begin{aligned} u' R_{ab} R_{bb}^{-1} R_{ba} u &= [\cdot1618 \quad \cdot3433 \quad \cdot2866] \begin{bmatrix} \cdot314 & \cdot229 \\ \cdot379 & \cdot258 \\ \cdot301 & \cdot294 \end{bmatrix} \begin{bmatrix} 1 \\ \cdot6 \end{bmatrix} \\ &= \cdot3931 \end{aligned}$$

and

$$u' R_{aa} u = [1 \quad .6] \begin{bmatrix} 1 & .638 \\ .638 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ .6 \end{bmatrix} \\ = 2.1256$$

and

$$r = \sqrt{\frac{.3931}{2.1256}} = .4300$$

We note that the battery of tests gives a better prediction of a complex criterion composed of assessments weighted in the ratio 1 : .6. We could find the maximum prediction in the Hotelling sense and the weights u and w associated with it (cf. Thomson, 6). For this purpose we require to solve the equation

$$\Delta = | R_{ab} R_{bb}^{-1} R_{ba} - \lambda R_{aa} | = 0$$

and we first calculate

$$R_{ab} R_{bb}^{-1} R_{ba} = \begin{bmatrix} .1181 & .2587 & .1587 \\ .0728 & .1410 & .2131 \end{bmatrix} \begin{bmatrix} .314 & .229 \\ .379 & .258 \\ .301 & .294 \end{bmatrix} = \begin{bmatrix} .1829 & .1404 \\ .1404 & .1157 \end{bmatrix}$$

$$\text{and} \quad \Delta = \begin{vmatrix} (.1829 - \lambda) & (.1404 - .638 \lambda) \\ (.1404 - .638 \lambda) & (.1157 - \lambda) \end{vmatrix} = 0$$

The larger root of this equation is $\lambda_1 = .18832$, from which $r_m = .4340$ and the weights u are in the ratio 1 : .27 and the weights w are respectively

$$.1377, \quad .2968, \quad .2162.$$

Computing maximum prediction in the sense defined by Hotelling is not so simple for batteries of assessments containing more than two assessments. One is required to solve quantics in λ of higher order than 2.

(b) *Calculating Maximum Battery Reliability.* By equation (11) we require to solve the equation

$$| R_{b\beta} - \lambda R_{bb} | = 0, \text{ for its largest root } \lambda_1.$$

Let us consider a battery of four tests, two practical ability tests (a and b) and two non-verbal paper tests (c and d). Then for these tests we have

$$R_{bb} = \begin{bmatrix} a & b & c & d \\ 1 & .518 & .390 & .288 \\ .518 & 1 & .372 & .109 \\ .390 & .372 & 1 & .402 \\ .288 & .109 & .402 & 1 \end{bmatrix}$$

The reliabilities of the tests are respectively,

$$.770, \quad .723, \quad .875, \quad \text{and} \quad .837.$$

Therefore by equation (6) we have

$$R_{b\beta} = \begin{bmatrix} .770 & .518 & .390 & .288 \\ .518 & .723 & .372 & .109 \\ .390 & .372 & .875 & .402 \\ .288 & .109 & .402 & .837 \end{bmatrix}$$

and so we have to solve

$$\Delta = | R_{b\beta} - \lambda R_{bb} | =$$

$$\begin{vmatrix} \cdot770 - \lambda & \cdot518(1 - \lambda) & \cdot390(1 - \lambda) & \cdot288(1 - \lambda) \\ \cdot518(1 - \lambda) & \cdot723 - \lambda & \cdot372(1 - \lambda) & \cdot109(1 - \lambda) \\ \cdot390(1 - \lambda) & \cdot372(1 - \lambda) & \cdot875 - \lambda & \cdot402(1 - \lambda) \\ \cdot288(1 - \lambda) & \cdot109(1 - \lambda) & \cdot402(1 - \lambda) & \cdot837 - \lambda \end{vmatrix} = 0$$

for its largest root.

At this stage a further advantage of the method becomes clear. It is very simple to calculate Δ when $\lambda=1$, for the non-diagonal entries disappear when λ is equal to unity and Δ becomes the product of the diagonal cells.

Thus
$$\Delta = (-\cdot230) \times (-\cdot277) \times (-\cdot125) \times (-\cdot163)$$

$$= + \cdot001298 \text{ approximately.}$$

It is to be noted that when $\lambda=1$, Δ is + ve for even orders and - ve for odd orders. Since we know that the largest root lies between $\cdot875$, the largest single test reliability, and unity, we try some value exceeding $\cdot875$, say, $\lambda=\cdot880$, and find that the corresponding value of Δ is $-\cdot00010959$, and, knowing the value and sign of Δ when $\lambda=1$, we are helped in our next guess. Starting from $\lambda=1$ we should proceed in the following stages :

λ	Δ
1	+ $\cdot00129809$
$\cdot880$	- $\cdot00010959$
$\cdot910$	- $\cdot00003082$
$\cdot917$	+ $\cdot00000411$

It is a good plan when three values of Δ have been obtained, at least one of which is - ve, to plot Δ against λ in order to find the approximate value of λ when $\Delta = 0$. If this is done $\Delta = 0$ corresponds roughly to $\lambda = \cdot9164$, and trying this value we obtain $\Delta = + \cdot00000074$.

The values of u which correspond to $\lambda = \cdot9164$ are obtained by calculating the elements of any row of the matrix $\text{adj} (R_{b\beta} - \cdot9164 R_{bb})$, that is, by suppressing any row of the matrix $(R_{b\beta} - \cdot9164 R_{bb})$ and calculating the minors formed from the remaining rows by suppressing each column in turn. If the fourth row is suppressed the values of the minors are found to be as follows :

$$\Delta_{41} = \cdot00045224$$

$$\Delta_{42} = \cdot00031358$$

$$\Delta_{43} = \cdot00112712$$

$$\Delta_{44} = \cdot00065956.$$

The weights $u_a u_b u_c u_d$ which give maximum reliability are therefore in the ratio approximately of

$$\cdot40 : \cdot28 : 1\cdot00 : \cdot59.$$

We can verify these weights by substituting in equation (10), by which

$$\rho = \frac{u' R_{b\beta} u}{u' R_{bb} u}$$

When $u' = \cdot40 \cdot28 \cdot1\cdot00 \cdot59$,

$\rho = \cdot9163$, which is in close agreement with the original figure of $\cdot9164$.

By Thomson's method the determinant $|R_{b\beta} R_{bb}^{-1} R_{b\beta} - \lambda R_{bb}|$ has to be formed. Upon calculation $R_{b\beta} R_{bb}^{-1} R_{b\beta}$ is found to be

$$\begin{bmatrix} .6202 & .4754 & .3833 & .2779 \\ .4754 & .5579 & .3599 & .1169 \\ .3833 & .3599 & .7720 & .3926 \\ .2779 & .1169 & .3926 & .7070 \end{bmatrix}$$

The determinant $|R_{b\beta} R_{bb}^{-1} R_{b\beta} - \lambda R_{bb}|$ is therefore equal to

$$\begin{vmatrix} .6202-\lambda & .4754-.518\lambda & .3833-.390\lambda & .2779-.288\lambda \\ .4754-.518\lambda & .5579-\lambda & .3599-.372\lambda & .1169-.109\lambda \\ .3833-.390\lambda & .3599-.372\lambda & .7720-\lambda & .3926-.402\lambda \\ .2779-.288\lambda & .1169-.109\lambda & .3926-.402\lambda & .7070-\lambda \end{vmatrix}$$

where λ is equal to r_m^2 . We may now substitute the numerical value of r_m^2 as found by the short method and the above determinant should vanish. The value of λ which we require is $(.9164)^2$, that is, .8398. On substituting this value and evaluating the determinant we find its value to be +.0000051. The minors formed by deleting the last row and each column in turn have the following values :

$$\Delta_{41} = .00171$$

$$\Delta_{42} = .00119$$

$$\Delta_{43} = .00427$$

$$\Delta_{44} = .00250$$

and the ratios of these values give the ratios of the test weights u . These ratios are as follows :

$$.40 \quad .28 \quad 1.00 \quad .59$$

and agree with the figures obtained by using the writer's method.

V. SUMMARY

1. The test weights w which give maximum prediction of an external complex criterion, formed from a number of assessments weighted arbitrarily by the vector of weights u , can be calculated by the equation

$$w' = u' R_{ab} R_{bb}^{-1}$$

where the assessments are the a variates and the predicting tests are the b variates.

The maximum correlation is given by

$$r = \sqrt{\left(\frac{u' R_{ab} R_{bb}^{-1} R_{ba} u}{u' R_{aa} u} \right)}$$

2. Maximum battery reliability of a battery of tests b is found by obtaining the maximum value of ρ where

$$\rho = \frac{u' R_{b\beta} u}{u' R_{bb} u}$$

Prediction of a Complex Criterion

This involves solving the equation

$$| R_{b\beta} - \lambda R_{bb} | = 0$$

for its largest root λ_1 . The root λ_1 is equal to ρ_m , the maximum reliability, and the weights w , which give maximum reliability, are in the ratio of the elements of any row of the matrix $\text{adj}(R_{b\beta} - \lambda_1 R_{bb})$.

The computations involved in these calculations are simple to carry out.

REFERENCES

1. Thomson, G. H. : *The factorial analysis of human ability*, 1946.
2. Hotelling, H. : 'The most predictable criterion.' *J. Educ. Psych.*, XXVI, 1935, 139-42.
3. Hotelling, H. : 'Relations between two sets of variates.' *Biometrika*, XXVIII, 1936, 321-77.
4. Thomson, G. H. : 'Weighting for battery reliability and prediction.' *Brit. J. Psych.*, XXX, 1940, 357-66.
5. Peel, E. A. : 'A short method for calculating maximum battery reliability.' *Nature*, CLIX, 1947, 816.
6. Thomson, G. H. : 'The maximum correlation of two weighted batteries.' *This Journal*, I, 1947, 27-34.

FACTOR ANALYSIS AND CANONICAL CORRELATIONS

By CYRIL BURT

Department of Psychology, University College, London

I. *The Case of a Single Factor.* II. *The Case of n Factors.* III. *Comparison of Canonical Factors and Principal Factors.* IV. *Algebraic Demonstration.* V. *Summary and Conclusions.*

I. THE CASE OF A SINGLE FACTOR

Aim. The purpose of this paper is twofold : (i) to indicate the type of psychological problem for which the calculation of 'canonical correlations' may usefully be employed ; and (ii) to determine the general relations between 'canonical analysis' and ordinary 'factor analysis.' In discussions of the subject, it appears that many psychologists doubt whether this particular technique has any value in psychology, and that the few who have considered it can commonly find little meaning for any canonical correlation other than the first. I hope to show that the procedure provides a more rigorous answer to questions which have often been raised in psychological researches and which have hitherto been treated by far more rough-and-ready means.¹

Illustrative Examples. One of the earliest problems to arise in the study of individual measurements was that of deciding whether the two general factors underlying two distinct sets of assessments formed in fact essentially one and the same general factor. How far, for instance, is the general factor that underlies a set of cognitive tests identical with the general factor that underlies two or more fallible estimates of the particular cognitive capacity that the tests are intended to measure ? Thus Spearman, in his earliest investigations of 'general intelligence' (1, 2), sought to show that 'general sensory discrimination' (as measured, for example, by tests of sound- and touch-discrimination) was identical with 'general intellectual ability' (as assessed, say, by two teachers, by two examinations, or by two scholastic tests, such as addition and completing a passage of prose).

Alternative Solutions. Various procedures have been suggested for dealing with such problems. The principles on which they are based can perhaps be most clearly exhibited if we apply them to one or two simple examples.

(1) Suppose, for instance, we have a set of correlations similar to those shown in Table I below. These, it will be observed, exhibit the 'hierarchical arrangement' as originally defined by Spearman.²

¹ Professor M. S. Bartlett, who has been good enough to read my paper in manuscript, points out that, whereas his article deals (p. 74) with the *differences* between factor analysis and canonical, mine has emphasized the possible *similarities*. For the psychologist, I think, the most important difference is that in canonical analysis the criterion is external (as in Spearman's earlier papers on intelligence (1, 2)), while in factor analysis the criterion is derived from the internal evidence of the tests or trait-measurements themselves (as in the later part of my own paper on intelligence-tests (6)). In an adequately planned research both types of criteria should be included. Then, when the criterion-assessments are multiple, and both a canonical analysis and a factor analysis are carried out, one finds in practice that the results are often surprisingly similar. It is with this special case, and the somewhat special type of conditions thus indicated, that my own discussion is chiefly concerned.

² The figures are in fact part of a table commonly used by Spearman to illustrate such an arrangement, and reproduced by one of his pupils (*Brit. J. Psych.*, VIII, 1916, p. 175).

TABLE I. HIERARCHICAL TABLE OF CORRELATIONS BETWEEN CRITERIA AND TESTS

Variables			(1)	(2)	(3)	(4)
Criteria	..	{ (1) (2)	—	.70	.60	.50
			.70	—	.50	.40
Tests	..	{ (3) (4)	.60	.50	—	.30
			.50	.40	.30	—

Spearman,¹ following a suggestion of Wissler's, proposed to take the ratio of (i) the average of the four cross correlations to (ii) the average of the two intercorrelations between the pair of criteria and the pair of tests. We thus obtain for the 'corrected correlation'

$$r_c = \frac{\frac{1}{4}(r_{13} + r_{14} + r_{23} + r_{24})}{\frac{1}{2}(r_{12} + r_{34})} \quad (\text{i})$$

$$= \frac{\frac{1}{4}(\cdot60 + \cdot50 + \cdot50 + \cdot40)}{\frac{1}{2}(\cdot70 + \cdot30)} = 1.00.$$

If (within the limits allowed by the probable error) the corrected correlation proved to be unity, then Spearman argued that the two general functions were identical.

(2) A more rigorous formula can be derived from the Pearson-Yule formula for partial correlation.² With this equation we are required to take, for both numerator and denominator, not the arithmetic mean of the observed correlations, but the geometric mean. This formula yields a 'corrected correlation' of unity only when the correlations form a matrix of rank one. For example, with Table II below we have

$$r_c = \frac{\sqrt[4]{(r_{13} \cdot r_{14} \cdot r_{23} \cdot r_{24})}}{\sqrt[2]{(r_{12} \cdot r_{34})}} \quad (\text{ii})$$

$$= \frac{\sqrt[4]{(\cdot63 \times \cdot54 \times \cdot56 \times \cdot48)}}{\sqrt[2]{(\cdot72 \times \cdot42)}} = 1.00.$$

Applied to a table of this type, Spearman's version with the arithmetic means (equation i) no longer gives a value of unity.

¹ Spearman and Krueger (2), p. 85. It should be remembered that Spearman at this stage conceived that, in an ideal 'hierarchical arrangement,' the correlations would descend by *equal* decrements in each row. Such a table would obey his 'intercolumnar criterion' perfectly. His accounts of the correction formula varied: in his earlier paper, he appears to have considered that, in theory at any rate, the most appropriate version would require the arithmetical mean in the numerator and the geometrical in the denominator (*Amer. J. Psych.*, XV, p. 90).

² Yule (4), p. 213; also given by Spearman.

TABLE II. CORRELATION MATRIX OF RANK ONE

Variables				(1)	(2)	(3)	(4)
Criteria	..	{	(1)	(.81)	.72	.63	.54
			(2)	.72	(.64)	.56	.48
Tests	..	{	(3)	.63	.56	(.49)	.42
			(4)	.54	.48	.42	(.36)

(3) In their usual form these equations are applicable only to tables involving *two* tests or 'predictors' and *two* reference-values or 'criteria.' If, however, in dealing with correlations such as those shown in Table II, we insert in the leading diagonal the self-correlations attributable to the single general factor only, we can derive a summation formula which will be applicable to any number of such variables and will yield the answer required. Let us designate the quadrants of the correlation matrix, partitioned as above, by

$$R = \begin{bmatrix} R_{aa} & R_{ab} \\ R_{ba} & R_{bb} \end{bmatrix}$$

and let us use ΣR_{ij} to designate the totals of the figures in the submatrix so indicated. Then the 'corrected correlation' will be

$$r_c = \frac{\Sigma R_{ab}}{\sqrt{\Sigma R_{aa}} \sqrt{\Sigma R_{bb}}} \quad (\text{iii})$$

Thus, with the figures given in Table II, we should have

$$r_c = \frac{.63 + .54 + .56 + .48}{\sqrt{(.81 + 2 \times .72 + .64)} \sqrt{(.49 + 2 \times .42 + .36)}} = 1.00.$$

The equation just given may be written

$$\frac{\Sigma R_{ba}}{\Sigma R_{aa}} = \lambda \frac{\Sigma R_{bb}}{\Sigma R_{ab}} \quad (\text{iv})$$

$$\text{where } \lambda = r_c^2 = \frac{\Sigma R_{ba} \Sigma R_{ab}}{\Sigma R_{aa} \Sigma R_{bb}}$$

$$\text{or } (\Sigma R_{ba}) (\Sigma R_{aa})^{-1} (\Sigma R_{ab}) - \lambda \Sigma R_{bb} = 0. \quad (\text{v})$$

Thus the formula for correction really depends on a proportionality between the sums of the correlations, analogous to the proportionality obtaining between the individual correlations in the case of a matrix of rank one.¹

¹ May I draw the attention of teachers of psychology to this somewhat obvious point? The analogies make it easier for the non-mathematical student to accept the elaborate form of the more generalized equation we shall reach below (equation (ix)). A similar structure, depending on a similar proportionality, may be noted in Pearson's formulæ for correcting for selection.

(4) If in the diagonal submatrices, R_{aa} and R_{bb} , we substitute units for the reduced self-correlations or 'communalities,' equation (iii) becomes the well-known equation for the correlation between the unweighted sum of the criteria and the unweighted sum of the tests. A formula based on this equation (modified to suit the assumption that the two 'pools' have the same total inter-correlation) was later adopted by Spearman for detecting or disproving the absence of 'specific correlation,' when 'reference-tests for g ' were used as the criteria (3, p. xxi). This, however, does not yield a correlation of unity; nor indeed does it furnish what Spearman originally required—the highest possible correlation between the 'pooled' criteria and the 'pooled' tests.

(5) But the correlation can evidently be increased, if we use weighted sums instead of unweighted; and it can be maximized by selecting both sets of weights appropriately. However, we are not always at liberty to weight our criteria. Thus, if our problem is to predict whether the pupils will or will not pass in a certain type of examination, we are not free to weight the component marks in that examination. If, on the other hand, we have *a priori* reasons for believing that the criteria (whether they be examination marks, teachers' estimates, or other psychological tests) do not as they stand yield the best possible criteria for the qualities we are seeking to determine, we may feel justified in giving differential weights to the criteria; and the only satisfactory data for determining the necessary weights may be the internal evidence, provided by criteria and predictors taken together.

Similar alternatives arise in attempting to verify the traditional doctrine of temperamental types. Our problem may be to determine how far a diagnosis based on body-measurements will agree with the psychiatrist's diagnosis of manic-depressive insanity or dementia præcox. In such a case there is no occasion for weighting the criteria. On the other hand, we may be interested in determining what particular pattern of temperamental tendencies can best be predicted by a suitable pattern of physical measurements. In that case we are dealing with what has been called 'bi-multiple correlation' (cf. 7, p. 153 and refs.); and the best available solution would seem to be obtained by Hotelling's formula for determining the 'most predictable criterion.'

Thus, if w_a and w_b denote the column-matrices of weights, we shall need to maximize

$$r_m = \frac{w_a' R_{ab} w_b}{\sqrt{(w_a' R_{aa} w_a) (w_b' R_{bb} w_b)}}, \quad (\text{vi})$$

or, to express the same requirement in an equivalent form, we shall need to find values for w_a and w_b , such that $w_a' R_{ab} w_b$ is stationary, subject to the condition that the variances $w_a' R_{aa} w_a = w_b' R_{bb} w_b = 1$. This leads (Kendall (5), p. 349) to

$$\left. \begin{aligned} w_a' R_{ab} - \mu w_b' R_{bb} &= 0 \\ w_b' R_{ba} - \nu w_a' R_{aa} &= 0 \end{aligned} \right\} \quad (\text{vii})$$

where μ and ν are undetermined multipliers. These reduce to the equation commonly cited, viz.,

$$(R_{ba} R_{aa}^{-1} R_{ab} - \lambda R_{bb}) w_b = 0 \quad (\text{viii})$$

where $\lambda = \mu^2 = \nu^2 = r_m^2$. As a condition for a non-trivial solution we have

$$| R_{ba} R_{aa}^{-1} R_{ab} - \lambda R_{bb} | = 0 \quad (\text{ix})$$

with analogous equations based on w_a and R_{bb}^{-1} . The reader will note the similarity in form between the expressions in equation (v) and in equations (viii) and (ix).

These formulæ will give one or more pairs of weight-vectors for calculating one or more 'canonical components' and the related values for the 'canonical correlations' (r_m). The question at once raised by this procedure is the following: what, if any, is the relation between the factors obtained by ordinary factor analysis and the factors which we may suppose to be responsible for these 'canonical correlations' and for Spearman's 'corrected correlation' (which, as we have seen, can be regarded as an approximation to the first canonical correlation)?

To answer this let us consider first the simplest possible case, that of a single common factor, as illustrated in Table II.

(a) *Canonical Analysis.* To determine the 'canonical components' we must insert units in the leading diagonal. λ is then found by inserting the given values for the correlations in equation (ix). We obtain $0.8236\lambda^2 - 0.4266\lambda = 0$, which gives $\lambda = r_m^2 = 0.5179$ or 0, and $r_m = 0.7197$ or 0. The proportional weights are then obtained by substituting the non-zero value for λ in equation (viii). The results are shown in the last column of Table III.

TABLE III. RESULTS OF FACTORIAL AND CANONICAL ANALYSIS

Variables			Factor Analysis		Canonical Analysis
			Saturations	Weights	Weights
Criteria	{ (1) (2)		.90	.5531	1.0000
			.80	.2595	.4689
Tests	{ (3) (4)		.70	.1603	1.0000
			.60	.1095	.6830

(b) *Factor Analysis.* On factorizing the correlations shown in Table II (with the squares of the saturations in the diagonal) we obtain the saturations and regressions (or 'weights') given in the first two columns of Table III. Using these weights to multiply the rows and columns of Table II (with units in the diagonal) we obtain

$$r_m = \frac{0.12546}{\sqrt{0.57991} \sqrt{0.05241}} = 0.7197, \text{ the same figure as before. Further, it will be}$$

noted that the weights for the two criteria and for the two tests have the same proportional values as the weights obtained by the canonical analysis.¹

II. THE CASE OF n FACTORS

Illustrative Example. Let us now consider the more complicated type of problem in which the correlations do not approximate to a matrix of rank one and are therefore no longer explicable in terms of a single general factor. We will take a case in which the number of criteria and tests is increased to $m=4$ and $n=4$. To keep the issues concrete, let us suppose that the 'predictors' are marks obtained in four tests, intended to measure intelligence (E, F, G, and H), and that the 'criteria' are estimates of intelligence provided by four teachers (A, B, C, and D). The correlations are shown in Table IV. As we shall see in a moment, they have been artificially constructed

¹ Applying the criterion implied by (xiii) (p. 105), we note $R_{ba}R_{aa}^{-1}R_{ab} = k_a R_{bb}^*$, where R_{bb}^* denotes the reduced correlation matrix (with factor variances in the diagonal), and k_a here = $\frac{r_{1g}^2(1-r_{2g}^2)+r_{2g}^2(1-r_{1g}^2)}{1-r_{1g}^2r_{2g}^2}$.

TABLE IV. OBSERVED CORRELATIONS BETWEEN TEACHERS' ESTIMATES AND TEST MEASUREMENTS

	Teachers				Tests			
	A	B	C	D	E	F	G	H
Teachers :								
A ..	1.0000	.5800	.4000	.1400	.4609	.8042	.5721	.5883
B ..	.5800	1.0000	.1000	-.0400	.5330	.1643	.8500	.5295
C ..	.4000	.1000	1.0000	.3000	.7814	.6333	-.3090	.3899
D ..	.1400	-.0400	.3000	1.0000	.5732	.5878	.0473	-.6065
Tests :								
E ..	.4609	.5330	.7814	.5732	1.0000	.6000	.2000	.2000
F ..	.8042	.1643	.6333	.5878	.6000	1.0000	.1600	.1700
G ..	.5721	.8500	-.3090	.0473	.2000	.1600	1.0000	.2400
H ..	.5883	.5295	.3899	-.6065	.2000	.1700	.2400	1.0000

to provide the simplest possible illustration of the argument ; but the general pattern has been suggested by an actual investigation in which C and D were ratings supplied by masters with an interest in the linguistic abilities of the pupils (teachers of English and French), and A and B by masters with a more practical interest (the Headmaster and the Science Master). E and F were verbal tests ; G and H non-verbal tests.

TABLE V. INVERSE OF THE CORRELATION MATRIX FOR THE CRITERIA (R_{aa}^{-1})

Variables			A	B	C	D
A ..	..	..	1.8460	-1.0153	-0.6012	-0.1188
B ..	..	..	-1.0153	1.5742	0.2056	0.1435
C ..	..	..	-0.6012	0.2056	1.3101	-0.3006
D ..	..	..	-0.1188	0.1435	-0.3006	-1.1126

(a) *Canonical Analysis.* Let us first calculate the canonical correlations by substituting in equation (ix) the requisite values from Table IV, and then find the standardized weights. The inverse of R_{aa}^{-1} is shown in Table V. On pre- and post-multiplying this inverse by R_{ba} and R_{ab} respectively we obtain the $n \times n$ matrix shown in Table VI. It will be seen that this is identical with R_{bb} . Accordingly, equation (ix) becomes $|I - \lambda I| \quad |R_{bb}| = 0$, so that we obtain n values for λ , all equal to 1.00.

TABLE VI. CALCULATED VALUE OF $R_{ba}R_{aa}^{-1}R_{ab}$

Variables			E	F	G	H
E ..	..	..	1.00	.60	.20	.20
F ..	..	..	.60	1.00	.16	.17
G ..	..	..	.20	.16	1.00	.24
H ..	..	..	.20	.17	.24	1.00

The weights can be calculated from equation (vii). In practice the simplest method is to proceed by successive approximation. We reach the figures in the last four columns of Table VII.

In the foregoing example, then, we have found that each of the canonical correlations between the 'predictors' and the 'criteria' (both being suitably and variously weighted) is perfect. The fact that the first of these correlations is 1.00 furnishes, for this set of data, the answer to Spearman's initial question: namely, how can we determine whether the 'general function' underlying a particular batch of tests, such as E, F, G, and H, is identical with the 'general function' underlying a set of teachers' estimates, or another set of tests used as 'reference-values.' Here, if we can regard the sum of the estimates A, B, C, and D, taken with positive weights (those shown in column ix), as indicating the boys' 'general intelligence,' we have demonstrated that it is possible, at least in theory, to procure a perfect correlation between intelligence as thus estimated and the combined set of tests.¹ What is still more interesting, if we may regard the differences between the estimates A and B on the one hand and the estimates C and D on the other as indicating the relative suitability of the boys for the Science or Arts side respectively, we can get an equally good prediction of this second characteristic by using the *same* battery of tests with a different set of weights. Here then we have a use for the second canonical correlation and its related weights.

The fact that in both cases we have to weight the criteria as well as the tests means that we not only check the relative validity of the several tests by the criteria, but also check the relative validity of the several criteria by the tests.²

(b) *Factor Analysis.* In his own researches, Spearman did not attempt to 'factorize' his tables, as we now understand factorization. He considered the procedure described in Section I—correlating pooled tests and pooled reference-values—to be more appropriate to his problems. Nowadays, however, the methods in common use follow the principle first adopted in my paper of 1909. This consisted in taking the observed correlation table as a whole, and finding what I called the 'highest common factor,' i.e., the factor which gave the closest fit to all the observed coefficients (6, pp. 161-2). Neither the procedure proposed by Spearman nor that followed in calculating 'canonical components' guarantees that the function or factor so obtained will be the factor which yields the *highest* correlation with the *separate* variables, or, what amounts to the same thing, will account for the largest proportion of the individual variation. To obtain such a factor as this we have to adapt the principles previously suggested by Pearson for calculating what he termed the 'lines of closest fit.' These lines, as he showed, are identical with the principal axes³ of the correlation ellipsoid (8); and the whole procedure is equivalent to the calculation of 'principal factors' by weighted summation.

¹ Note that, by this procedure, i.e., by selecting the appropriate weights, we can (with suitable data) secure a perfect correlation between tests and estimates, even when both involve *more than one common factor*. It may be remembered that Spearman himself doubted this possibility; and maintained that it would be essential to use a 'purified' battery of tests, yielding (as near as possible) a perfect hierarchy with no supplementary factors.

² The need for this step emerged early in correlational work, as soon, in fact, as moderately efficient tests became available for general or special abilities (6, p. 158). Many teachers' estimates of intelligence are apt to be biased by overrating the importance of the pupils' mechanical memory. Similarly, many teachers' estimates of a special capacity, like verbal ability, are biased by including the influence of general intelligence as well.

³ Pearson's procedure was based on the method of least squares, and led to the familiar equation (expressed in matrix form) $(R - \lambda I)w = 0$: (cf. his equations (ix) and (x), *loc. cit.*, p. 562). An alternative proof, yielding the same result, may be based on the requirement that each variable shall be weighted in proportion to its correlation with the general factor (as illustrated in my previous article in this *Journal*, I, p. 15). Since $MM'w = R w$, this leads at once to $R w = \lambda w$. As teachers of statistical psychology have often to deal with students whose mathematical equipment falls short of the calculus, may I suggest this simple deduction as a convenient teaching-device?

TABLE VII. SATURATIONS AND WEIGHTS FOR PRINCIPAL AND CANONICAL FACTORS

Variable	Factor Analysis								Canonical Analysis			
	Saturations				Weights				Weights			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
A ..	.8802	-.2095	-.0147	-.4256	.2420	-.0954	-.0103	-.5774	.4862	-.1775	-.0219	-.12561
B ..	.6928	-.5773	.2600	.3452	.1905	-.2630	.1818	.4683	.3827	-.4891	.3879	1.0188
C ..	.6530	.4771	-.5549	.1950	.1795	.2174	-.3880	.2646	.3607	.4042	-.8279	.5756
D ..	.3596	.7587	.5427	.0225	-.0989	.3457	.3794	.0305	.1986	.6427	.8097	-.0665
Square Sum ..	1.8104	1.1804	.6703	.3389	.1368	.2450	.3276	.6236	.5524	.8472	1.4920	2.9514
E ..	.8238	.3418	.0139	.4521	.2265	.1557	.0097	.6134	.4508	.3369	.0183	1.1352
F ..	.7990	.4107	-.0022	.4392	.2197	.1871	-.0015	-.5959	.4373	.4048	-.0029	-.1.1028
G ..	.5013	-.6118	.6113	-.0274	.1378	-.2787	.4274	-.0372	.2744	-.6030	.8043	-.0688
H ..	.5088	-.5956	-.6214	-.0153	.1399	-.2714	-.4345	-.0208	.2785	-.5871	-.8176	-.0384
Square Sum ..	1.8272	1.0145	.7600	.3983	.1381	.2106	.3716	.7331	.5473	.9857	1.3157	2.5110
Total Square Sum ..	3.6376	2.1949	1.4303	.7372	.2749	.4556	.6992	1.3567	1.0997	1.8329	2.8077	5.4624
Column ..	i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii

Let us then factorize our entire correlation table in this way (keeping units¹ in the leading diagonal). The resulting factor saturations and regressions are shown in Table VII.

III. COMPARISON OF CANONICAL FACTORS AND PRINCIPAL FACTORS

What, then, are the relations between the two sets of factors? Factors we may consider to be identifiable in terms of their weights: (for this principle see 7, p. 161). Accordingly, let us examine the first factor to begin with; and compare the weights assigned by the two modes of analysis to the four *tests*: (cols. v and ix, last 4 rows). Those found by the canonical method turn out to be exactly proportional to those found by the factor analysis of the whole table: they are in fact 1.991 times the latter. Hence, so far as the predictors are concerned, the factor responsible for the first canonical correlation is here identical with the first factor obtained by an ordinary factor analysis. The difference in their absolute amounts is due merely to the difference in the number of variables included, and the requirement that the factor-measurements shall in either case be in standard measure. Turning next to the weights assigned to the criteria (i.e., the teachers' estimates), we again note a similar proportionality: the weights reached by the canonical method are 2.009 times those obtained by factorization.

Now, in factorizing the table as a whole, the first factor obtained is a general factor common *both* to the teachers' estimates *and* the tests. The first canonical factor, therefore, is in the present case *identical* with the first factor obtained by factorizing the combined table in the usual way. Similarly, on taking the three remaining factors in turn, we shall find that in each case the canonical factor is identical with the corresponding factor obtained by factor analysis.

So exact an agreement will, of course, not be obtained in practice. What are the conditions, then, which will make, or tend to make, this agreement more or less complete? If we factorize *separately* first the correlations between the teachers' estimates (R_{aa}) and then the correlations between the tests (R_{bb}), we shall find that (provided we use weighted summation, not simple) the saturations obtained for both factors from these 4×4 tables are identical with those obtained from the whole 8×8 table. Thus in the present instance the highest common factor found in the teachers' estimates and the highest common factor found in the tests are each identical with the highest common factor obtained from the whole eight variables treated as a single battery. Things which are identical with the same thing must be identical with one another. Hence, each of the two factors obtained from the smaller symmetrical submatrices must be identical with the other.

¹ If the reader wishes to know what would be the outcome of a comparison based on 'communalities' (instead of units) in the leading diagonal, he will have no difficulty in carrying out a canonical analysis and a factor analysis with the values for R_{aa} and R_{bb} given in Table IV above. To secure perfect canonical correlations, however, R_{ab} will now have to be modified to fit the cross products of the new factor saturations. Limits of space preclude an extended discussion of this further type of case here; but a detailed calculation will be found at the close of the roneo'd "Laboratory Notes on Factor Analysis and Canonical Correlations" (obtainable from the laboratory) on which this paper has largely been based. I should like to acknowledge the help received from my assistant, Dr. Banks, in carrying out part of the calculations and in checking calculations of my own.

This result depends not so much on the submatrices of intercorrelations as on the submatrix of cross-correlations. It is in fact essential that these cross-correlations between the tests on the one hand and the teachers' estimates on the other should be analysable in terms of the factor saturations obtainable from each set of intercorrelations taken separately. In geometrical terms the requisite condition amounts to this: for the canonical correlations to be perfect, the vectors representing the tests must lie in the same space as the vectors representing the criteria.¹

IV. ALGEBRAIC DEMONSTRATION

These conclusions can easily be demonstrated algebraically. Adopting the notation suggested above, let us suppose that the rank of R_{aa} (or R_{bb} whichever is the smaller) is the same as its order. A correlation matrix can always be factorized by weighted summation so that $R=FF'=LVL'$, where L is an orthogonal matrix of direction-cosines, and V is the diagonal matrix of factor variances. With this analysis the factor-saturations are given by $F=LV^{\frac{1}{2}}$, and the weights by $W'=V^{-\frac{1}{2}}L'=V^{-1}F'$. We thus have

$$M=FP, \quad (x)$$

where M denotes the matrix of observed test measurements and P the uncorrelated factor measurements, duly standardized so that $PP'=I$ (7, and refs.).

Moreover, just as we can factorize the whole matrix R in this way, so we can factorize each of the submatrices R_{aa} and R_{bb} . We then obtain $R_{aa}=L_aV_aL'_a$ and $R_{bb}=L_bV_bL'_b$. A simple and sufficient condition for all the canonical correlations to be perfect is accordingly:

$$R_{ab}=F_aF'_b=L_aV_a^{\frac{1}{2}}V_b^{\frac{1}{2}}L'_b, \quad (xi)$$

for then the expression in the determinantal equation (ix) becomes

$$\begin{aligned} & R_{ba}R_{aa}^{-1}R_{ab}-\lambda R_{bb} \\ &= (L_bV_b^{\frac{1}{2}}V_a^{\frac{1}{2}}L'_a)(L_aV_a^{-1}L'_a)(L_aV_a^{\frac{1}{2}}V_b^{\frac{1}{2}}L'_b)-\lambda R_{bb} \\ &= L_bV_bL'_b-\lambda R_{bb} \\ &= (I-\lambda I)R_{bb} \end{aligned}$$

so that $\lambda_1=\lambda_2=\dots=\lambda_n=1.00$. Conversely, if all the latent roots are unity, we have $R_{ba}R_{aa}^{-1}R_{ab}=R_{bb}$. Hence $R_{ab}=F_aHF'_b$, where H may be any orthogonal matrix,² and has here been taken to be I . Under this condition it is easy to see that on factorizing the whole matrix we obtain

$$F'=[F'_a, F'_b] \text{ and } V=[V_a+V_b].$$

Thus, if the first canonical correlation is to be 1.00, the ideal experimental procedure would be to choose the variables in the two sets—criteria and predictors respectively—so that the first factor in the one is identical with the first factor in the other, and consequently with the first factor in the whole group; and similarly for the supplementary factors.

¹ This implies (as we shall see in a moment) a certain arbitrariness in choosing the weights in the ideal or artificial case. In practice the canonical correlations will never be absolutely perfect; hence this freedom of choice then disappears.

² Substituting from (x) for $R_{ba}=M_bM'_a$ etc., we have, after elimination, $P_bP'_aP_aP'_b=I$: so that $P_bP'_a=H_{ab}$ an orthogonal matrix. Further, if $H_{ab}=I$, so that $R_{ab}=F_aF'_b$, we may still write $R_{ab}=F_aH_x.H'_x F'_b$, where H_x may be any orthogonal matrix.

Again, since by equation (vii) $w'_a R_{ab} - \lambda w'_b R_{bb} = 0$, we have, when $\lambda = 1$, $w'_a F_a H = w'_b F_b$; and this will be satisfied if we take w'_a equal to a column of $F_a^{-1} = V_a^{-\frac{1}{2}} L_a$, and w'_b equal to the corresponding column of $F_b^{-1} = V_b^{-\frac{1}{2}} L_b$. With these weights the supplementary condition is also satisfied, viz., that $w'_a R_{aa} w_a = w'_b R_{bb} w_b = 1$.

When examining an empirical correlation table, comprising two sets of variables, the requirements expressed in equation (xi) may be formulated in more concrete terms. It is simply that the appropriate *sections* of the columns of saturations, obtained by factorizing the whole table, must be 'uncorrelated'¹ with one another. Or, in symbols, if $L' = [L_a, L_b]$, then $L_a L'_a = I$ and $L_b L'_b = I$. For this, the simplest test in practice is that, when the submatrix of cross-correlations is multiplied by its transpose, the resulting product-matrix should yield the same proportionate factors as the diagonal submatrix of the same order as the product-matrix: i.e., $F'_a (R_{ab} R_{ba}) = V_a V_b F'_a$ or $F'_b (R_{ba} R_{ab}) = V_b V_a F'_b$. However, this test does not imply a necessary condition.

When $R_{ab} = F_a H_a \cdot H'_b F'_b$, the factors obtained from the submatrices R_{aa} and R_{bb} , like the factors implied by the canonical analysis, will no longer be identical with those reached by factorizing the entire table. If, however, the positive and negative halves of the successive bipolar factors are about equally represented and about equally balanced in the two sets of variables (in the 'tests' and in the 'criteria'), then the canonical factors will approximate more or less closely to those obtained by factorizing the whole table. This broader condition is very frequently fulfilled in empirical correlation-matrices with which the psychologist has to deal, particularly when his investigations have been so planned that his 'criteria' relate to the same essential qualities as his 'predictors' are intended to estimate.² When, as in the case of factors for body-build and temperamental assessments, each set of variables has its own general factor which overlaps scarcely at all with the general factor in the other set, then the first pair of canonical factors may correspond with the second pair of principal factors (i.e., the first bipolar factors) obtained from the separate sets.

If (as in Table II) the intercorrelations form a matrix of rank one, F will consist of a single column only. Then $R_{ba} R_{aa}^{-1} R_{ab} = f'_b f'_a R_{aa}^{-1} f_a f'_b$, where $f'_a R_{aa}^{-1} f_a$ is evidently a scalar, k_a say. Since $f'_b w_b$ will also be a scalar, c say, equation (viii) reduces to

$$w_b = \kappa R_{bb}^{-1} f_b \quad (\text{xii})$$

where $\kappa = \frac{1}{\lambda} c k_a$. Thus, w_b is proportional to $R_{bb}^{-1} f_b$, the weights obtained by factorizing R_{bb}^* in the ordinary way. It must therefore be proportional to the weights for the single general factor obtained from the *whole* correlation matrix. Thus, with a hierarchical correlation table, the two components—canonical and factorial—will still be identical, even though the canonical correlation, r_m , is no longer unity.

In fact, since $c = f'_b w_b = \frac{1}{\lambda} c k_a \cdot f'_b R_{bb}^{-1} f_b$, we have

$$\lambda = r_m^2 = f'_a R_{aa}^{-1} f_a \cdot f'_b R_{bb}^{-1} f_b = k_a \cdot k_b = \frac{S_a}{1+S_a} \cdot \frac{S_b}{1+S_b}, \quad (\text{xiii})$$

where S denotes the familiar expression $\Sigma \frac{r_{ig}^2}{1-r_{ig}^2}$.

¹ By this is meant that the 'unadjusted correlation' or product-sum must be zero.

² A special case of some interest arises when (as in some of the enquiries alluded to above) the 'criteria' consist, not of continuously graded variables, but of a dichotomous or multiple classification into discontinuous categories, yielding a point-distribution rather than a normal distribution. Since $\Sigma X\bar{X} = \Sigma \bar{X}X$, we then have $R_{ab} = R_{ba} = R_{aa}$; and hence equation (ix) reduces to $|R_{ba} - \lambda R_{bb}| = 0$: so that (viii) specifies a kind of multiple 'discriminant function,' which may be conveniently employed to determine the differences between *more than two* groups (cf. 7, p. 153).

V. SUMMARY AND CONCLUSIONS

1. When we have two sets of variables relating to one and the same quality or qualities (e.g., predictive tests and criterion estimates for the same mental abilities, or again physical and mental symptoms of the same temperamental types), various methods of analysis may be adopted according to the precise nature of the problem raised. We may, for example, factorize the two sets together as though they formed a single battery ; or we may factorize each set independently, and correlate homologous factors ; or finally we may determine the maximum correlation between the two sets of variables (or their residuals), each set being combined to form a single composite variate. The question discussed is the relation between the results of these various modes of analysis.

2. When the estimates supplied for the criterion are fixed and final, then the problem is simply one of weighting the predictors to yield the best correlations with those criteria, taken severally or together. But, when the estimates supplied are themselves as fallible or speculative as the predictors, the method of canonical correlation, which weights the criteria as well as the tests, is more appropriate. In either case an instructive check is obtained by factorizing the two sets together.

3. In practice, with empirical data, the factors obtained by the different procedures will not be exactly identical. But cases arise in which they reveal considerable similarity.

4. In theory, even when the correlations between the predictors involve more than one factor, it may still be possible to obtain a maximum correlation of 1.00 between a weighted combination of the predictors and a weighted combination of the criteria. Indeed, the same set of predictors, combined with different weights, may even yield perfect predictions for different criteria, or for the same criteria differently weighted.

5. The necessary and sufficient condition for such 'canonical correlations' to be 1.00 is that the factors obtained by analysing the two sets as a single battery are identical with the factors obtained by analysing the separate sets, or with linear combinations derived from those factors by a set of orthogonal weights.

REFERENCES

1. Spearman, C. : 'General intelligence objectively determined and measured.' *Amer. J. Psych.*, III, 1904, 253-285.
2. Spearman, C., and Krueger, F. : 'Die Korrelation zwischen verschiedenen geistigen Leistungen.' *Zeitsch. f. Psych.*, XLIV, 1906, 60-85.
3. Spearman, C. : *Abilities of Man*. London : Macmillan & Co., 1926.
4. Yule, G. U. : *Introduction to the Theory of Statistics* (2nd ed.). London : Griffin & Co., 1912.
5. Kendall, M. G. : *The Advanced Theory of Statistics*, II. London : Griffin & Co., 1946.
6. Burt, C. : 'Experimental tests of general intelligence.' *Brit. J. Psych.*, III, 1909, 94-177.
7. Burt, C. : 'The unit hierarchy and its properties.' *Psychometrika*, III, 1938, 151-168.
8. Pearson, K. : 'The lines of closest fit to a system of points.' *Phil. Mag.*, II, 1901, 559-572.

FLYING ABILITY AND BODY-BUILD

By CHARLOTTE BANKS

Psychological Department, University College, London

I. *Problem.* II. *Assessments of Body-build.* III. *Assessments of Flying Ability.*
IV. *Correlations between Body-build and Flying Ability.* V. *Summary and Conclusions.*

I. PROBLEM

The relation between physical and mental types has formed a long-standing problem in individual psychology. The application of a factorial technique, however, appears to have established the essential facts: their explanation remains a more speculative and complex question. First, there seems undoubtedly to be a small but significant correlation between the physical characteristics supposed to be distinctive of the traditional temperamental types and the mental or emotional characteristics usually attributed to them (1). But secondly, so far as existing methods of assessment go, the correlations obtained are much too low for the prediction of psychological qualities from body-build to be of any practical value (2).

Nevertheless, during the early years of the war, the belief that physique might be used as an index of temperament received a strong impetus from the work of Dr. Sheldon and his colleagues in America (3); and in this country, owing partly to the widespread popularity of Kretschmer's theories among medical psychologists, and partly to Sheldon's own researches for the American Air Force, the suggestion was frequently put forward that observations of the physical type of recruits might be of service as a guide to personnel selection, particularly in regard to air-crew.

As an extension of his earlier work, Dr. Sheldon, in a series of unpublished memoranda, based on material collected at Randolph Field (Texas), claimed that efficient air pilots could be selected, prior to training, on the basis of body-build. He used his well-known devices for somatotyping (4); and obtained data which, so he believed, demonstrated that 'mesomorphs' were much more likely to make good pilots than 'endomorphs' or 'ectomorphs.' The question naturally arose whether some such method of pilot selection should not at least be given a trial in this country; and the Director of Training Research, Air Ministry, referred Dr. Sheldon's memoranda to Sir Cyril Burt for comment and advice.

The first step was to obtain more adequate evidence for the existence and nature of physical types as traditionally conceived. Burt had suggested that this problem could best be solved by means of factor analysis. In earlier work with school children he had obtained definite evidence for such types on a fairly large scale (5, p. 643), and had confirmed this with smaller groups of adult students (2, 6). Evidence from larger numbers, however, was urgently needed. Thanks to the interest of Air Marshal Sir Harold Whittingham and the generous co-operation of Dr. Morant, it proved possible to secure accurate measurements for nine physical traits for 2,400 Royal Air Force recruits. These were divided into eight age-groups, and the correlations factorized by simple summation (8, 9). With all the groups the results obtained were fully significant and highly consistent. They revealed a clear bipolar factor, contributing nearly 15 per cent. to the total variance, and dividing both traits and persons into leptomorphic and pachymorphic groups respectively. These two groups correspond approximately with what Sheldon terms 'ectomorphs' and 'endomorphs.' Accordingly, we need have no hesitation in accepting the possibility of some such classification into two main antithetical types according to body-build.

The relation of these types to temperamental characteristics and to operational efficiency is a more difficult problem. As is well known, the leptomorphic type is supposed to show a marked tendency towards introversion, and the pachymorphic towards extraversion. It was accordingly suggested that the pachymorphs might conceivably yield the best type of recruit for fighters and for pilots engaged on short, sharp, intensive operations, and that the leptomorphs might yield the best type for bombers, navigators, long-distance flights, instrumental flying, and the like. Others argued that the best pilots would be those with a well-balanced temperament, neither markedly introverted nor markedly extraverted. If there were any element of truth in these speculations, we might expect to find some correlation between body-build, on the one hand, and assessments of flying ability, on the other.

We are here confronted by a type of statistical problem which has long been familiar in work on individual psychology. We have two sets of data for comparison, (a) a number of diagnostic assessments (physical measurements in the present case), and (b) a number of independent assessments of the criterion (marks or measurements for efficiency in various types of air-crew duty or more generally for temperamental characteristics relevant to such work): our aim is to determine the relation between the two. As Burt has pointed out, there are two alternative lines of approach which are not always sufficiently distinguished—an empirical and a theoretical.

(i) We might endeavour, by 'weighting,' 'correcting,' or otherwise 'improving' both sets of measurements, to find out by empirical means what particular combination of each yields the highest discoverable correlation. Thus, using a weighted linear combination of physical measurements, on the one hand, and another for flying performances, on the other, we might seek to ascertain, not merely (as in the ordinary regression problem) what particular combination of diagnostic traits will yield the best predictions, but also what particular combination of the independent estimates will yield (in Hotelling's phrase) 'the most predictable criterion.' We should then, no doubt, go on to seek the apparent reasons for this correlation by inspecting the nature of those particular traits or of those particular criteria which had received the highest weights. In investigating the relations between physical and temperamental assessments, Burt has used this procedure both with children and with adults; but he points out that the maximized correlation, so attained, appears to be a confused resultant of many obscure causal linkages—differences in health, nutrition, endocrine balance, social conditions, as well as in temperamental type or tendency in the traditional sense. He argues, therefore, that, as in dealing with intelligence and cognitive tests, so in such problems as these: it will generally be better to start with a theoretical assumption about the possible relation between this or that well-defined physical factor and this or that well-defined mental type of operational performance, and to investigate the connexion between the two (7).

(ii) The second approach therefore proceeds by first defining what may be called a diagnostic type (or a diagnostic quality) on the basis of internal evidence alone, i.e., regardless of the external evidence of the criteria. We have then to investigate the relation between this definite type or quality and the practical performances or activities in which we are interested. Our task, in fact, is not to predict what is most easily predicted, but to predict the subsequent performances as they are actually reported. In the problem before us, therefore, we should begin by defining a physical type on the basis of the factor analysis of body-measurements already described. The first bipolar factor, determined in the preliminary research, may be taken as representing, in precise, quantitative form, the tendency towards one or other of the two main types (the ecto- or leptomorphs and the endo- or pachymorphs respectively); and, with the regression equation so obtained, we can calculate this bipolar factor for each individual recruit.

This then was the procedure adopted in the present enquiry. A preliminary trial was made with somewhat limited data, collected for a different purpose. Three physical measurements only were available. With these earlier data it appeared that (as had already been suspected) the factor-measurements for body-build produced, not two or three sharply distinguishable types, but a continuous distribution, conforming fairly closely with the normal curve. So far as the accessible evidence indicated, there seemed to be no appreciable correlations between body-build and either allocation¹ to air-crew duty (as bomber, pilot, or navigator) or general flying

¹ Some of the earlier data submitted to us suggested that men of heavier and broader physique might be more suited for the larger and heavier types of aircraft. However, the present paper will not be concerned with problems of allocation.

efficiency. However, to reach reliable conclusions it was felt that fuller and more accurate information was needed (7).

II. ASSESSMENTS OF BODY-BUILD

In the earlier investigation with 2,400 recruits it was found that, taking the first bipolar factor (based on all nine measurements) as the criterion, a very good multiple correlation could be secured by using four specially selected measurements. The best combination appeared to be Sitting Height, Leg Length, Weight, and Waist Circumference. Using the factor-saturations obtained from the nine measurements and the partial regressions obtained from these four, the multiple correlation, calculated in the usual way (8), proved to be as high as 0.96.

Dr. Morant has accordingly very kindly obtained for us (a) measurements for these four physical traits and (b) records of flying assessments for 134 pilots from Elementary and Service Flying Training Schools. The physical traits were measured by Dr. Morant personally in accordance with the procedure described in his R.A.F. Memorandum (10). For purposes of the present enquiry they have been reduced to standard measure. Table I gives the partial regression coefficients used for weighting these four measurements to procure a single assessment of body-build. The figures were derived from the correlations calculated for the foregoing 2,400 men.

TABLE I. PARTIAL REGRESSIONS FOR BODY-TYPE FACTOR

Trait			Sitting Height	Leg Length	Weight	Waist
Regression	..	..	0.3224	0.8731	-0.8765	-0.2140

Table II gives the frequency distribution of the factor-measurements for body type, calculated with these weighting coefficients. A negative factor-measurement indicates a pachymorphic tendency, and a positive factor-measurement a leptomorphic. In the last column the frequencies obtained in the previous enquiry have been added for comparison. These were based on 805 cases, but only three body-measurements were available (Standing Height, Weight, and Leg Length). In the present study the type-measurements were based on four body-measurements; and these referred to a slightly different combination of traits (Sitting Height, Weight, Leg Length, and Waist Circumference). In spite of these differences the resulting distributions are closely similar.

The percentages are plotted in the form of a graph in Fig. 1. The distribution evidently approximates closely to the normal, but is slightly skewed at the pachymorphic end. The asymmetry is presumably due to the inclusion of Weight as a dominant item for the pachymorphic tendency. However, chi-squared is only 3.56, and $P=0.73$. Consequently, the deviation from normality is too slight to be statistically significant.

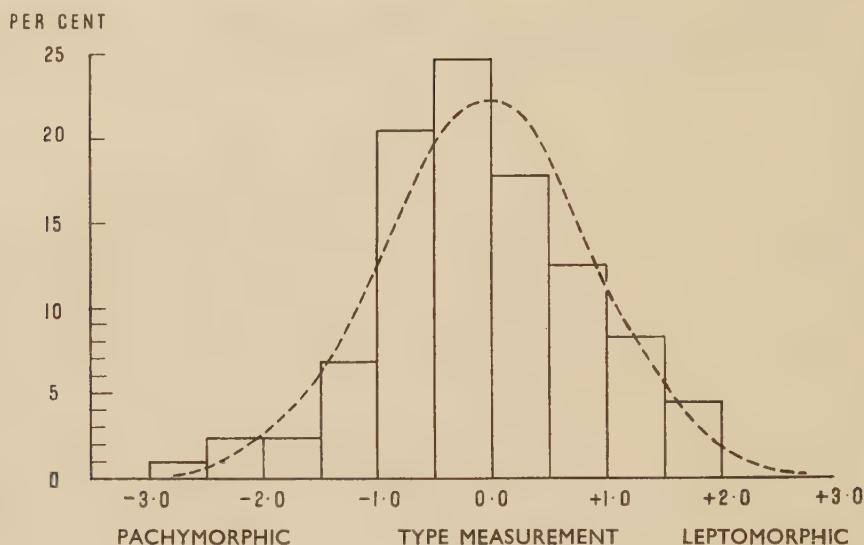


FIG. 1.

TABLE II. DISTRIBUTION OF FACTOR-MEASUREMENTS FOR BODY-TYPE FACTOR

Factor-measurement	(a) Based on 4 Body-measurements			(b) Based on 3 Body-measurements
	Number	Observed Frequency Per cent.	Normal Frequency Per cent.	Observed Frequency Per cent.
-3.0 to -2.5 ..	1	0.7	0.3	0.4
-2.5 to -2.0 ..	3	2.2	1.1	1.4
-2.0 to -1.5 ..	3	2.2	4.3	4.1
-1.5 to -1.0 ..	9	6.7	9.5	9.7
-1.0 to -0.5 ..	27	20.2	16.7	17.0
-0.5 to 0.0 ..	33	24.6	21.9	20.1
0.0 to 0.5 ..	24	17.9	20.6	18.3
0.5 to 1.0 ..	18	13.6	14.3	13.9
1.0 to 1.5 ..	10	7.5	7.8	7.7
1.5 to 2.0 ..	6	4.4	2.7	3.9
2.0 to 2.5 ..	—	—	0.7	1.8
2.5 to 3.0 ..	—	—	0.1	1.1
3.0 to 3.5 ..	—	—	—	0.6
Total	134	100.0	100.0	100.0

III. ASSESSMENTS OF FLYING ABILITY

All the 134 men were assessed for flying efficiency, first at the Elementary and later at the Service Flying Training Schools. For the figures and other relevant information I am especially indebted to Dr. J. B. Parry, Head of Science 4, Air Ministry. For the methods of grading, see (11). At both schools the marks allotted are divided into three categories—'Ground Total,' 'General Flying,' and 'Total Flying.'

'Ground Total' consists of the total marks obtained at examinations on work in five theoretical subjects: airmanship, armaments, meteorology, navigation, and signals. Of these the first two receive the highest weight. 'Ground Total' may thus be regarded as assessing knowledge of the theory of flying. 'General Flying' includes practical efficiency in all basic air work—e.g., take off and climb, medium turns, steep turns, gliding turns, approach, landing, general alertness, style, handling, and co-ordination. 'Total Flying' includes the marks for this basic work together with additional marks for more specialized work, such as instrument flying, night flying, and Link trainer. To the total so obtained General Flying contributes about half.

Table III gives product-moment inter-correlations within schools and between schools.

TABLE III. CORRELATIONS BETWEEN MARKS FOR FLYING ABILITY

<i>Between Schools</i>					Correlations	
Elementary with Service : Ground Total	..	..	..	..	0.43	
Elementary with Service : General Flying	..	..	..	..	0.34	
Elementary with Service : Total Flying	..	..	..	..	0.47	
<i>Within Schools</i>					Elementary	Service
Ground Total and General Flying	..	..	..	..	-0.06	-0.08
Total Flying and Ground Total	..	..	..	..	0.33	0.32
Total Flying and General Flying	..	..	..	..	0.52	0.50

Judged by a comparison of the marks obtained at the first and second flying schools, the 'reliability' of the assessments is of the order of .30 to .50. Within the two schools the correlations are remarkably consistent. In both cases, General Flying has an insignificant negative correlation with theoretical knowledge as indicated by Ground Total; Total Flying correlates with General Flying to the extent of about 0.50, and with the written examinations by about 0.30.

IV. CORRELATIONS BETWEEN BODY-BUILD AND FLYING ABILITY

In comparing flying success with body type, as here assessed, there are two main questions to be borne in mind. First, does one or other of the two main types—leptomorphic or pachymorphic—make the more efficient flyer, so far as can be ascertained by the marks awarded? Secondly (as Sheldon's data seem rather to imply), does the average or intermediate type do better than either of the two extremes?

In order to answer the first question, product-moment correlations were calculated between the factor-measurements and the various flying assessments. The coefficients obtained are given in Table IV below. A positive coefficient indicates a positive correlation between a leptomorphic tendency and success in flying. The standard errors are approximately ± 0.09 . All the correlations are therefore insignificant. Accordingly, we may safely conclude that a tendency towards leptomorphic or pachymorphic physique is of no value whatever for predicting probable success in flying.

TABLE IV. CORRELATIONS BETWEEN LEPTOMORPHIC BODY-BUILD AND FLYING SUCCESS

Mark for Flying					Elementary School	Service School
Ground Total	..	..	..	..	0.04	-0.06
General Flying	..	..	..	..	0.03	0.03
Total Flying..	..	..	..	..	0.02	0.09

In order to test the second hypothesis, that definite pachymorphic and leptomorphic types make less successful flyers than the intermediate type, various devices were tried. The simplest seemed to be to divide the entire group into two classes—(a) those having factor-measurements between -1.0 and $+1.0$, and (b) those having factor-measurements beyond these limits. These two subgroups contained 32 and 102 individuals respectively. Tetrachoric correlations were then calculated between presence or absence of a marked body type as thus assessed and achievement below or above the average in the six flying assessments. These correlations are shown in Table V, and probably give a fair indication of the relations in question.¹

TABLE V. CORRELATIONS BETWEEN AVERAGE BODY-BUILD AND FLYING SUCCESS

Mark for Flying					Elementary School	Service School
Ground Total	..	..	..	..	-0.33	-0.03
General Flying	..	..	..	..	0.19	0.30
Total Flying..	..	..	..	..	0.06	-0.02

As regards theoretical work, men of average body-build apparently do not succeed at the Elementary School so well as those of a definite physical type (whether pachymorphic or leptomorphic); but the tendency is slight, and is not maintained at the Service School. As regards General Flying, however, both at the Elementary and at the later stages, there is a clear tendency for the intermediate or average type to gain higher marks. These latter correlations thus in some measure bear out Dr. Sheldon's conclusions. However, the two tendencies work in opposite directions, and thus appear to neutralize one another in the total marks. In both the Elementary and the Service Schools the correlation between Total Flying and average body-build is virtually zero.

It is tempting to suggest that the men of intermediate or normal physique are naturally likely to be the more efficient in outdoor tasks demanding physical activity or prompt and accurate motor co-ordination, and to be more at ease in manipulating the controls in a cockpit of standard size. On the other hand, those of marked leptomorphic or pachymorphic type might be expected to include persons of more

¹ The correlation ratios are (a) at the Elementary School, -0.22 , 0.34 , and 0.15 , (b) at the Service School, -0.19 , 0.34 , and -0.09 . The use of the tetrachoric correlation can only be defended as a rough and approximate estimate; but the data so far available hardly warrant a detailed discussion of the actual distributions. It is to be hoped that a fuller investigation may be attempted at a later stage.

sedentary habits, perhaps a few mild pathological cases, and possibly several with a latent psychoneurotic instability. However, the frequency-distribution indicates that all cases of a markedly abnormal type, whether leptomorphic or pachymorphic, had already been excluded. All such explanations, therefore, must be regarded as highly speculative; and any definite interpretation of the figures must await further investigations on the relation between body-build and other personal characteristics.

V. SUMMARY AND CONCLUSIONS

1. Measurements and assessments have been obtained for 134 Air Force recruits in regard to (a) certain bodily traits and (b) efficiency in the theory and practice of flying. By means of partial regression coefficients, factor-measurements have been calculated for the bipolar factor responsible for the distinction between the so-called leptomorphic and pachymorphic types. These measurements have then been compared with the marks obtained from the Elementary and Service Flying Training Schools.

2. The factor-measurements fully confirm Burt's contention that what are called 'physical types' are in fact merely more extreme instances from either tail of a perfectly continuous distribution, which in form approximates very closely to that of the normal curve. The distribution here obtained with four body-measurements (Sitting Height, Waist, Leg Length, and Weight) agrees remarkably well with that previously obtained from a different group of men with three slightly different body-measurements (Standing Height, Leg Length, and Weight only). The close correspondence thus indicates that the factor in question is fairly stable from one type of enquiry to the other.

3. No correlation was found at any stage between the tendency towards a leptomorphic type (or the reverse) and flying efficiency, whether assessed by theoretical knowledge or by practical performance.

4. In practical flying the man of intermediate or average body-build does slightly better. In theoretical work, on the other hand, his achievements, at least in the earlier stages, are apparently not quite so good.

5. The correlations are too low to be of practical value for personnel selection; but suggest problems of interest to the theoretical psychologist for further research.

REFERENCES

1. Burt, C.: 'The mental differences between individuals.' *Brit. Ass. Ann. Rep.*, 1923, pp. 215-239.
2. Burt, C.: 'Analysis of temperament.' *Brit. J. Med. Psych.*, XVII, 1938, pp. 158-188.
3. Sheldon, W. H., Stevens, S. S., and Tucker, W. B.: *The Varieties of Human Physique*. Harper Bros., 1940.
4. Sheldon, W. H.: *Physique and Temperament*. Harper Bros., 1942.
5. Burt, C.: *The Backward Child*, Univ. of London Press, 1935.
6. Burt, C.: 'The factorial study of physical types.' *Man*, XLIV, 1944, pp. 82-6.
7. Burt, C.: 'Relation between body-build, temperamental characteristics, and flying efficiency.' R.A.F. (T.R.2) Memo., 1945 (unpublished).
8. Banks, C.: 'Factor analysis applied to data from H.M. Forces.' Doctorate Thesis (1945). Univ. of London Library.
9. Burt, C., and Banks, C.: 'A factor analysis of body-measurements for British adult males.' *Ann. Eug.*, XIII, 1947, pp. 238-256.
10. Morant, G.: *Report on a Survey of Body and Clothing Measurements of Royal Air Force Personnel*. Flying Personnel Research Committee, Report No. 633 (a), 1945.
11. Air Ministry: *Grading: A Report on Pilot Selection in the Royal Air Force*. Air Ministry, 1945 (restricted).

THE PRIMARY PERSONALITY FACTORS IN WOMEN COMPARED WITH THOSE IN MEN

By RAYMOND B. CATTELL

University of Illinois

I. *Scope of the Enquiry.* II. *Precautions for Valid Ratings.* III. *The Subjects and the Variables.* IV. *The Factor Analysis.* V. *Description of the Personality Factors.* VI. *Summary. References.*

I. SCOPE OF THE ENQUIRY

The loading patterns of some fourteen personality factors—those of largest variance—in adult men have been published previously² and ten of them have been tolerably confirmed and clarified⁵ apart from whatever further confirmation results from scattered studies limited to one or two factors at a time.⁴

Consequently the next objectives are (1) to discover how these primary personality factors modify their patterns with such important circumstances as age, sex, cultural status, etc., and (2) to shift the measurement of the factors from ratings or questionnaires to objective tests. This article is an initial attempt to map sex modification in the light of a study of 240 women, chosen to match in age and education the 133 men of the previous study.

Certain features in the plan of this research are due to its being aimed also at the second objective above: the bridging of factor studies in three distinct areas of measurement—(1) Behaviour ratings in natural life situations—or “B R data”; (2) Questionnaire and self-rating—or “Q data”; and (3) Objective tests—or “T data.” For to test the hypothesis that the factors found in these three realms of observation are actually the same unitary traits outcropping in different fields of expression it has been necessary to subject this same population to the somewhat long and exacting experiment required to make possible all required inter-correlations. The verdict as to the relationship of the present B R factors to objective test or T factors and the ‘mental interiors’ viewed in such Q factors as those of Guilford⁸ will be available in articles now in the press.⁶

II. PRECAUTIONS FOR VALID RATINGS

The emphasis which the writer has placed in this series of researches upon the desirability of basing the initial definition of the factors on B R data needs some explanation. For in some circles it is uncritically assumed that ratings are unreliable, subjective, and lacking in power to discriminate variables. When carried out hastily, by persons without time or taste for the subtleties of behaviour, this is demonstrably true; but it must be insisted also, first, that under scientifically specified conditions ratings are highly valid, and secondly, that they have certain advantages over other data which make their omission at certain stages of investigation a grave research blunder.

The more substantial of these advantages are: (1) *Perspective on factor space.* If a research begins with laboratory tests or questionnaires, some extremely narrow specific segment of total real life behaviour may appear as a general factor, while major aspects of personality may be unknowingly omitted. “It is the scale that

creates the phenomenon," as a well-known physicist and philosopher of science, Charles Guye, has well said. The concept of the *personality sphere*,⁴ of a truly representative choice of variables relative to the total behaviour of man that is important to man, has to rest on semantics and on a rating approach. (2) The establishment of a factor in terms of well-known, everyday traits generally points the way to the general nature of the objective tests likely to provide the desired more rigid, objective measures of it. The converse step, from a laboratory test factor to the areas in which it is likely to be important in everyday life, is less easy to take. (3) The behaviour expressive of important factors of personality is usually a complex *pattern* of action, such as the majority of tests yet proposed, counting simple, isolated reactions, fail to catch. The only instrument yet available which can proceed at once to make the delicate perception, abstraction, and estimation of a whole response pattern is the human mind.

In most conditions of rating these advantages are never attained because insufficient trouble or control is exercised to gain common reliability. The following minimum conditions discussed earlier^{4 5} were observed in this and the preceding study. (1) The raters lived in the same house with the ratees. (2) The latter were peers of the ratees, standing in no official relationship to them. (3) The raters were asked to rate no more than sixteen people, on the assumption that a large circle could not be well known. (4) They rated all subjects with respect to one trait (at any given time), not one subject with respect to all traits. (5) They were not asked to make finer than a five-point rating—actually, average, above average, and below, with a later choice of the extreme case from the above and below average groups to constitute the first and fifth point score. See⁵ for detail. (6) They were drilled in the behavioural definition and illustration of the traits. (7) The ratings of fifteen people (the people living in the same sorority house) were pooled to give the trait scores of each individual. (8) The raters were given a month to study the ratees after the definition of the trait behaviour, though they had already lived with them six months or more. (9) The raters were adequately motivated by talks on the scientific value of the project and by payment for their time.

Reliabilities were worked out for twelve of the traits, taken at random, by pooling the scores of the first half (eight) of the raters and correlating against the pool of the second half. Since it has been shown by Burt, Carter, Guilford, and others that extra raters operate precisely like extra test items in their effect upon reliability, the "full length" reliability (for the pool of the whole group, as used in our factorizing) was calculated by the Spearman-Brown formula. The coefficients for these twelve traits (numbered here as in the following section) were: (1) .86; (3) .87; (6) .90; (7) .81; (8) .83; (9) .83; (10) .89; (20) .88; (21) .85; (26) .85; (28) .74; (30) .76. Variable, (36) which was not a rating but an actual intelligence test, had a reliability in this group of .82.

It should be noted that these reliabilities are not, as in the first of our three studies,² worked out from the correlations within each small group, appropriately averaged for all groups, but for the whole group of 240 cases at once. The shift from working correlations on initial *rankings* of persons within each small group to *numerical ratings* thrown into a single large group was deliberately undertaken in this and the second study to see how far the results of factorization would be independent of such a marked change of rating method.

III. THE SUBJECTS AND THE VARIABLES

The subjects were 240 undergraduate women students in an American State university, living in sororities and independent houses, averaging about 20.7 years

of age and being random with respect to academic field. Thus they represent a group selected with respect to education and social status from among women of their age, but not excessively so, since in this State approximately 15% of the school population goes on to the University and this fraction does not correspond with the most intelligent 15% of the population.

As Burt¹ has aptly emphasized, proper representation of variables is as important as proper sampling of the population. In the initial research to determine the main factor space for describing personality we took steps to get an unbiased sample of all personality variables, as indicated in the concept of the personality sphere. Since our next purpose has been to sharpen and clarify the nature of factors already brought into some degree of focus, we proceeded here with the following guiding principles:—

(1) To represent each known factor by a minimum of two variables, chosen from the six most highly saturated in that factor in the previous study.

(2) To represent each factor by a variable which, from previous studies of all kinds,³ seemed most likely to be the essence of the factor. That is, we attempted to find a purer measure, with higher saturations than any previously found. This is part of an iterative, “distillation” procedure, in which each research profits by the preceding to get closer to the nature of the factor.

(3) To re-shape the definition of landmark variables slightly from their original “logical trait” form toward “organic trait unity” form,*⁴ in the light of associates discovered for them in the previous factorization.

That the objectives of (1), (2), and (3) were to some extent achieved is shown (since there is no increase of reliability of ratings *per se* in the parallel research relative to the original study to account for greater inter-correlation) by the fact that the communality for most variables was definitely increased in the two present studies.

(4) Where factors of a similar or co-operative⁴ nature existed in the previous study, e.g., the two schizoid factors A and H; the three character factors C, G, and J, we proceeded to choose variables that would “pull them apart” better, i.e., bring out the finer characteristics by which they differ.

(5) To give relatively more variables to factors needing sharper definition, particularly the factors (H and onwards) of smaller variance (in the original study), was our last aim. Thus four variables were inserted to bring out factor K, but only three for the well-known dominance factor E and only one—an actual intelligence test—to mark the “intelligence-in-personality” factor B.

The defined variable list actually given to the raters deliberately dispersed these factor groupings and reversed the polarity of items indiscriminately, as shown below.

In the author’s M.S. each variable is described at some length. Lack of space has necessitated the reduction of the list to the titles only, except in the case of the first two, retained as illustrations.

*The variables with which an experimenter starts are normally defined by arbitrary, logical limits existing in the mind of the experimenter. In time the form of the unitary trait as it exists in nature emerges. The experimenter can then measure variables in the same area, defined anew better to align themselves with the true, organic nature of the functional unity. The contrast of “logical” and “organic” trait unities can be readily illustrated by comparing the trait unities of the phrenologist with those found in later studies of brain function, or by the modern factor analysts. (See ⁴, p. 126.)

TABLE I. DEFINITIONS OF VARIABLES *

1. <i>Readiness to Co-operate</i>	vs. <i>Obstructiveness</i>
Generally tends to say yes when invited to co-operate. Outgoing. Ready to meet people at least half-way. Finds ways of co-operating despite difficulties.	Inclined to raise objections to a project, cynical or realistic. "Cannot be done." Uninterested or unfavourable attitude to joining in. Inclined to be "difficult."
2. <i>Emotionally Stable</i>	vs. <i>Changeable</i>
Can be depended upon to look at questions objectively, without emotional prejudice, and in the same constant light from day to day. Above emotion in his judgments. Dependable and realistic.	Sees things in terms of the emotion of the moment. Emotional bias changes from day to day and place to place. Does not remain the same person from day to day. Undependable.
3. <i>Attention-getting</i>	vs. <i>Self-sufficient</i>
4. <i>Assertive, Self-assured</i>	vs. <i>Submissive</i>
5. <i>Depressed, Solemn</i>	vs. <i>Cheerful</i>
6. <i>Frivolous</i>	vs. <i>Responsible</i>
7. <i>Attentive to People</i>	vs. <i>Cool, Aloof.</i>
8. <i>Easily Upset</i>	vs. <i>Unshakable Poise, Tough.</i>
9. <i>Languid, Slow</i>	vs. <i>Energetic, Alert.</i>
10. <i>Boorish</i>	vs. <i>Intellectual, Cultured</i>
11. <i>Suspicious</i>	vs. <i>Trustful.</i>
12. <i>Good-natured, Easygoing</i>	vs. <i>Spiteful, Grasping, Critical.</i>
13. <i>Calm, Phlegmatic</i>	vs. <i>Emotional.</i>
14. <i>Hypochondriacal</i>	vs. <i>Not So.</i>
15. <i>Mild, Self-effacing</i>	vs. <i>Self-willed, Egotistic.</i>
16. <i>Silent, Introspective</i>	vs. <i>Talkative.</i>
17. <i>Persevering, Determined</i>	vs. <i>Quitting, Fickle.</i>
18. <i>Cautious, Retiring, Timid</i>	vs. <i>Adventurous, Bold.</i>
19. <i>Hard, Stern</i>	vs. <i>Kindly, Soft-hearted.</i>
20. <i>Insistently Orderly</i>	vs. <i>Relaxed, Indolent.</i>
21. <i>Polished</i>	vs. <i>Clumsy, Awkward.</i>
22. <i>Prone to Jealousy</i>	vs. <i>Not Prone to Jealousy.</i>
23. <i>Rigid</i>	vs. <i>Adaptable.</i>
24. <i>Demanding, Impatient</i>	vs. <i>Emotionally Mature.</i>
25. <i>Unconventional, Eccentric</i>	vs. <i>Conventional.</i>
26. <i>Placid</i>	vs. <i>Worrying, Anxious.</i>
27. <i>Conscientious</i>	vs. <i>Somewhat Unscrupulous.</i>
28. <i>Composed</i>	vs. <i>Shy, Bashful.</i>
29. <i>Sensitively Imaginative</i>	vs. <i>Practical, Logical.</i>
30. <i>Neurotic Fatigue</i>	vs. <i>Absence of Neurotic Fatigue.</i>
31. <i>Æsthetically Fastidious</i>	vs. <i>Lacking Artistic Feeling.</i>

* The items of behaviour included under one definition are put together partly on logical grounds, but also, and predominantly, on the evidence of previous research on the constituents of correlation clusters or "surface traits." Thus, in trait 20, miserliness and orderliness are included, though not logically related, and, in trait 13, proneness to emotional display of both joy and sorrow are together, because of previous evidence of the existence of a trait of "general emotionality." (See ⁴, Chapter 8, for this evidence.)

TABLE II. PERSONALITY TRAIT CORRELATION MATRIX

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 ..																	
2 ..	.449																
3 ..	-.234	-.568															
4 ..	-.098	-.124	-.633														
5 ..	-.380	-.033	-.222	-.098													
6 ..	-.568	-.633	-.450	-.048	-.124												
7 ..	-.652	-.236	-.073	-.151	-.566	-.269											
8 ..	-.008	-.394	-.103	-.498	-.061	-.206	.114										
9 ..	-.359	-.113	-.302	-.426	-.677	-.058	-.389	.198									
10 ..	-.344	-.489	-.211	-.184	-.053	-.583	-.222	-.359	.180								
11 ..	-.545	-.533	-.339	-.222	-.501	-.255	-.502	-.124	-.323	.246							
12 ..	-.601	-.289	-.294	-.429	-.455	-.230	-.670	-.215	-.181	-.115	-.518						
13 ..	-.115	-.581	-.555	-.308	-.274	-.341	-.108	-.379	-.330	-.218	-.281	.092					
14 ..	-.232	-.516	-.384	-.228	-.256	-.226	-.125	-.175	-.175	-.150	-.505	-.243	-.396				
15 ..	-.536	-.440	-.603	-.658	-.145	-.397	-.492	-.313	-.113	-.174	-.526	-.694	-.341	-.327			
16 ..	-.133	-.308	-.582	-.435	-.629	-.266	-.412	-.013	-.624	-.155	-.067	-.148	-.597	-.139	.275		
17 ..	-.414	-.504	-.232	-.211	-.169	-.683	-.040	-.302	-.186	-.541	-.154	-.017	-.263	-.207	.112	.217	
18 ..	-.120	-.141	-.509	-.423	-.675	-.300	-.272	-.208	-.685	-.059	-.201	-.145	-.416	.070	.256	.745	.253
19 ..	-.517	-.114	-.315	-.575	-.297	-.211	-.614	-.490	-.047	.074	.432	-.723	.040	.058	-.714	.006	.097
20 ..	-.067	-.295	-.269	-.070	-.399	-.513	-.171	-.107	-.113	-.272	.196	-.300	.272	.003	-.021	.355	.625
21 ..	-.399	-.422	-.169	-.100	-.065	-.443	-.367	-.272	-.155	-.555	-.304	.174	.272	-.052	.201	.143	.346
22 ..	-.486	-.550	-.533	-.463	-.247	-.350	-.424	-.000	-.047	-.249	-.677	-.633	-.379	.401	-.697	-.186	-.110
23 ..	-.398	-.185	-.145	-.274	-.580	-.009	-.492	-.061	-.309	-.084	.585	-.634	-.019	.315	-.440	.249	.230
24 ..	-.588	-.550	-.589	-.498	-.159	-.507	-.496	-.022	-.021	-.283	-.565	-.644	-.412	.375	-.742	-.215	-.240
25 ..	-.161	-.153	-.448	-.385	-.133	-.283	-.115	-.342	-.276	-.023	-.087	-.133	-.125	-.017	-.431	-.315	-.104
26 ..	-.030	-.454	-.167	-.111	-.076	-.066	-.018	-.586	-.087	-.079	-.356	-.025	-.526	-.537	-.028	.126	.126
27 ..	-.487	-.486	-.610	-.370	-.075	-.644	-.356	-.184	-.110	-.425	-.123	.399	.324	-.182	.651	.352	.443
28 ..	-.228	-.079	-.469	-.637	-.423	-.024	-.323	-.488	-.645	-.248	-.197	-.054	-.222	.029	-.300	-.614	.102
29 ..	-.126	-.244	-.179	-.107	-.309	-.200	-.301	.319	-.156	-.056	-.144	-.288	-.341	.148	.132	-.227	-.275
30 ..	-.249	-.490	-.399	-.223	-.225	-.216	-.145	-.257	-.115	-.099	.542	-.271	-.485	.636	-.303	-.152	-.179
31 ..	-.035	-.041	-.018	-.098	-.111	-.064	-.104	-.144	-.003	-.172	.056	-.181	.111	.066	-.111	.197	.198
32 ..	-.050	-.249	-.486	-.273	-.325	-.354	-.118	-.039	-.329	-.248	-.070	.020	-.364	.150	-.302	-.423	-.259
33 ..	-.081	-.120	.504	.516	-.443	-.142	-.271	-.121	-.526	.059	-.016	.001	-.463	.138	-.292	.738	-.077
34 ..	.215	-.267	.465	.245	-.668	.268	.457	.160	-.540	.233	-.150	.269	-.505	.095	-.087	.780	-.325
35 ..	-.237	-.459	-.010	-.511	.019	.458	-.151	.608	.333	.704	.152	.082	-.160	.142	.130	.042	-.488
36 ..	-.032	-.084	-.082	-.119	-.102	.003	-.087	-.248	-.005	-.135	-.040	-.136	.056	-.071	-.103	.043	.145
Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

32. *Marked Interest in Opposite Sex* vs. *Slight Interest in Opposite Sex*.
 33. *Frank, Expressive* vs. *Secretive, Reserved*.
 34. *Gregarious, Sociable* vs. *Self-contained*.
 35. *Dependent, Immature* vs. *Independent-minded*.

To the above variables we added an estimate of 36. *Intelligence* vs. *Deficiency of Intelligence* obtained by use of the A.C.E. test in the normal course of college entrance. This one objective test was common to both the behaviour rating factorization and the objective test factorization.

IV. THE FACTOR ANALYSIS

Using the product-moment coefficient the experimenter first calculated the correlation matrix shown in Table II.* It was factored by the centroid method, yielding the unrotated factor matrix shown in Table III. The communalities were

* For each person the score on each trait was derived by addition of the scores assigned by fifteen voters each on a five-point scale (see page 115 above). The possible score thus ranged from $15 \times 1 = 15$ to $15 \times 5 = 75$. The product-moment correlations used these final scores as the variates.

estimated by taking the highest correlation in each column, a procedure which, on the basis of this and other experience, the writer is convinced leads to overestimation. (Better results have since been obtained with the method recommended by Medland.) Certainly the obtained communalities in Table III err on the side of being too large. Tucker's criterion indicated that extraction was complete after the tenth or possibly eleventh factor.[†] It was stopped after the eleventh on the assumption that rotation would remove the eleventh if superfluous.

† The absolute sums of the residuals after the extraction of the 10th, 11th, and 12th factors were respectively 15.53, 14.53, 14.27. The expression $\{(\sum s + 1)/\sum s\}^{\frac{1}{2}}$, when s is 10, equals 0.967 and when s is 11 equals 0.991. The criterion value $(n-1)/n$ equals 0.972; so that by Tucker's method the 11th is the last factor, though the nearness of 0.967 to 0.972 suggests that even the 10th might be the last.

TABLE III. UNROTATED FACTOR MATRIX F

Variable	Factor												h ²
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	
1 ..	-.48	-.47	.33	.36	-.06	-.10	.03	.10	.11	.21	-.07	-.16	.801
2 ..	-.75	-.03	.41	-.08	.11	.08	.09	-.13	-.05	-.04	.06	.09	.797
3 ..	.80	.06	.27	.15	-.12	.02	-.04	.14	.13	.09	-.09	.06	.812
4 ..	.47	.34	.65	.21	.09	.10	-.03	-.05	.02	.08	-.14	-.05	.854
5 ..	-.17	.74	-.40	.01	-.03	.19	.07	-.02	.12	.05	.03	-.02	.797
6 ..	.70	-.09	-.25	-.39	-.22	-.09	.03	.10	.04	-.09	.04	.04	.793
7 ..	-.18	-.74	.19	.31	.14	.32	-.17	.25	.18	-.17	.22	-.16	1.061
8 ..	.12	-.38	-.73	.28	.15	-.19	.08	.19	-.28	.15	-.05	.09	.983
9 ..	-.22	.40	-.67	-.13	-.10	.16	.22	.16	-.04	.07	-.13	.06	.811
10 ..	.47	-.19	-.43	-.33	.23	-.16	.35	.14	.20	.14	-.08	-.13	.854
11 ..	.48	.57	-.37	.22	.17	.17	.08	-.03	-.17	-.07	.12	-.09	.862
12 ..	-.40	-.75	.02	.09	-.19	.11	-.04	.07	.07	.07	-.04	.03	.798
13 ..	-.70	.26	.09	-.38	-.06	.05	.16	.11	.07	-.09	.08	.08	.780
14 ..	.46	.21	-.27	.45	-.20	.17	.13	-.05	.13	.05	-.05	-.09	.649
15 ..	-.70	-.52	-.26	.08	-.05	.06	.06	.04	.05	.04	.10	.14	.879
16 ..	-.62	.51	-.40	-.17	-.13	-.07	.07	.16	-.06	-.03	-.04	-.10	.902
17 ..	-.50	.31	.40	.33	.27	-.14	-.17	-.10	.12	.18	.14	.02	.813
18 ..	-.51	.45	-.58	.16	.13	.05	.12	.16	.07	-.10	-.11	.11	.923
19 ..	.38	.66	.33	-.35	.20	.02	.06	-.16	.04	.07	.08	-.08	.900
20 ..	-.37	.51	.03	.33	.38	-.17	.10	-.07	.06	-.11	.12	-.06	.729
21 ..	-.40	.11	.36	.36	-.53	-.42	.31	-.35	-.19	-.09	-.13	.17	1.197
22 ..	.65	.49	-.05	.14	.14	-.12	-.04	.08	.07	-.06	.05	.05	.740
23 ..	.15	.72	-.16	.22	.20	.06	-.08	.02	.03	-.04	.04	.13	.686
24 ..	.74	.44	-.03	-.03	.08	-.12	-.04	.05	.03	.09	-.07	.14	.801
25 ..	.40	.07	.39	-.24	-.22	.24	-.07	.17	-.11	.19	.21	-.07	.612
26 ..	-.25	.06	.48	-.62	.02	.11	.10	.09	-.02	.02	.13	-.04	.731
27 ..	-.72	-.13	-.16	.44	.17	.08	-.04	-.17	-.12	-.13	-.05	.12	.869
28 ..	.29	-.16	.78	.18	-.13	.14	.11	-.08	.14	-.13	-.12	-.03	.857
29 ..	.19	-.37	-.07	.18	-.38	-.05	-.23	.18	-.16	.16	.07	.10	.509
30 ..	.49	.21	-.27	.44	-.18	.12	.02	-.10	.04	.18	.11	-.04	.656
31 ..	-.10	.28	.15	.23	-.27	-.40	.26	.11	-.12	-.15	.22	-.18	.594
32 ..	.46	-.23	.27	.04	-.22	-.22	.20	.07	.19	-.19	-.11	.03	.566
33 ..	.51	-.28	.45	.13	.20	.20	.04	-.11	-.04	.13	.06	.09	.682
34 ..	.51	-.62	.28	.20	.05	-.09	.12	-.02	.06	-.07	.08	.08	.810
35 ..	.23	-.26	-.68	-.20	.05	-.20	.22	.11	.17	.13	.10	.08	.788
36 ..	-.02	.17	.15	-.10	.13	-.10	-.29	-.33	.37	.12	.26	.18	.533

It will be noticed that the communalities for variables 7 and 21 ran slightly above unity. Such theoretically impossible values are not unknown in centroid analyses, being due to poor estimates of residual communality on these variables in relation to the variance left after each factor has been abstracted.

The discovering of simple structure, always a difficult undertaking with as many as twelve factors, proved more difficult than in any previous study undertaken by the writer. After 23 rotations of all factors (253 individual axis rotations) only seven factors had attained simple structure. Examination of their psychological meaning showed them to be factors A, B, C, F, G, K, and something resembling I. (Note the similarity to the factors which steadied early in the men's study.⁵) In view of the satisfactoriness of the experimental basis of this study—the lower P.E.'s of the correlations and the excellent reliability coefficients (see above section), running

nearly 20 points higher than those for the men's studies, the difficulty over simple structure was puzzling. It was decided to check the whole basis of the factorization—the extraction of factors, the original correlations, and the pooling of the ratings—but no error except of the most trivial kind was found. After ten more total rotations it was decided to extract a twelfth factor, assuming that we were trying to obtain simple structure on too few factors. Simple structure was finally obtained for all but two factors on the 43rd rotation, though for final polishing and the settling of the last two factors the process had to be carried to the 54th over-all rotation.

It is interesting to note that in the attempt to retain tolerably orthogonal angles with respect to a factor that appears to have a sure hyperplane one sometimes loses other factors that have really 'settled' and then picks up these same simple structures again—often numbered as a different factor—later. The fact that these positions are picked up, as it were, from different directions at different times gives one, however, greater confidence in their reliability. Some delay always results because initially easily found hyperplanes slough off as residuals and one has to start out again looking for hyperplanes belonging to factors of substance. Again, here as in previous rotations we isolated factors which, as their hyperplanes improved, came closer and closer in angle, eventually proving to be one and the same factor.

Apart from the possibly frivolous explanation that our difficulties arose from the comparative subtlety of the feminine mind, no really adequate explanation suggests itself for this prolonged labour on rotation extending to 588 single rotations and requiring eighteen months of full-time work after the factor extraction. (The grouping method of factor extraction favoured the brevity of the first twelve-factor study rotation² but also gave us apparently a false factor—D.) Possibly the slight inflation of many loadings due to the persistent tendency of the guessed communalities to run too high is responsible for the difficulty.

But the writer feels that the attainment of a truly sound simple structure is so essential to obtaining meaningful, psychologically transactable source traits that one should be fully prepared to regard rotation as the major part of a factor analysis, as far as the expenditure of time and skill is concerned.

V. DESCRIPTION OF THE PERSONALITY FACTORS

The results of rotation are shown in Table IV. In accordance with the writer's previous practice the figures represent the correlation of the variables with the reference vectors—which for each vector are simply proportional to the loadings on the corresponding factor. For at this stage of research, not involving any specification equation, the labour of working out the conversion ratios is unnecessary. The rotation matrix and the direction numbers of vectors one to another are shown in Table V.

We shall now discuss the factors individually in approximate order of their contribution to the variance of all variables. Where the factor clearly resembles one already named,^{4 5} it is given the same identification letter and name: otherwise labelling is carried further down the alphabet. The nature of a factor is defined by the variables it does not affect at all and by those it affects most highly, usually

the highest six, depending upon whether a natural break can be made there. Numbers on the left represent number of the variable (Table I) and its correlation with the reference vector, the polarity of the written description of the variable being adjusted to the sign of the correlation, so that all names on the left are positively loaded with the factor.

TABLE IV. ROTATED MATRIX (REFERENCE VECTOR STRUCTURE) $V = FA$

Variable		Factor											
		1	2	3	4	5	6	7	8	9	10	11	
1	..	.00	.55	.45	.13	-.03	.01	-.02	.02	-.05	-.01	-.02	
2	..	.46	.08	.24	-.02	.04	.06	.05	.08	.09	.06	-.20	
3	..	-.16	.03	.02	.28	.00	-.01	.12	.03	-.06	-.12	.28	
4	..	.05	-.12	.25	.33	.01	.06	.01	.05	-.10	-.00	.19	
5	..	-.08	-.13	.09	.22	-.34	-.19	-.10	-.02	.13	.06	-.01	
6	..	-.06	-.23	-.48	-.08	.04	-.06	.10	.01	-.01	-.11	.07	
7	..	-.06	.43	.14	.20	.34	-.04	.23	-.57	.04	-.05	-.06	
8	..	-.49	.13	-.02	-.40	.15	-.06	-.12	.10	-.23	-.20	.04	
9	..	-.05	-.06	-.07	.03	-.40	-.25	-.13	.08	-.13	-.05	.02	
10	..	.04	-.13	.02	-.06	.03	-.56	.07	.02	-.13	-.02	.06	
11	..	-.37	-.43	-.04	.05	.12	-.06	-.03	-.06	-.02	-.10	-.04	
12	..	-.03	.50	.00	.06	-.01	.01	.02	-.11	-.01	.04	-.14	
13	..	.56	.03	.05	.00	-.22	-.06	.11	.08	.08	-.06	-.03	
14	..	-.52	.00	.06	.40	-.16	-.08	.08	.01	.03	.05	-.07	
15	..	.02	.42	.10	-.09	.02	-.07	.06	-.04	.12	.01	-.23	
16	..	.12	-.02	-.05	-.09	-.44	.03	-.13	.10	-.06	-.06	.08	
17	..	.10	.25	.57	.05	.05	.09	-.04	.03	.26	.04	.09	
18	..	.01	.03	.21	-.04	-.33	-.19	.01	-.02	-.09	.07	.11	
19	..	.29	-.46	.05	.11	.05	-.14	-.06	.07	.12	.00	.05	
20	..	.08	-.12	.47	-.06	.06	-.01	.14	.05	.11	.06	.02	
21	..	-.01	.03	.06	-.06	-.25	.57	.24	.78	.04	.06	-.26	
22	..	-.16	-.26	.05	.02	.07	-.02	.12	.01	.02	-.09	.27	
23	..	-.11	-.20	.19	.06	-.06	-.04	-.01	-.04	.07	.00	.19	
24	..	-.11	-.25	-.01	.00	.00	-.08	-.02	.11	-.03	-.08	.30	
25	..	.04	.01	-.19	.25	.10	.10	-.05	-.02	.08	-.37	.04	
26	..	.65	-.10	-.07	.03	.03	-.07	.03	.01	.04	-.17	-.03	
27	..	-.10	.15	.24	-.14	.04	.11	-.04	-.05	-.01	.22	-.21	
28	..	.19	.02	.14	.40	.05	.10	.30	.03	-.06	.08	-.02	
29	..	-.40	.36	-.24	-.04	.02	.32	-.09	.06	.01	-.27	.09	
30	..	-.62	.03	.04	.30	-.01	-.00	-.02	.05	.17	-.06	-.11	
31	..	-.05	-.05	.04	-.06	-.04	.41	.34	.42	.06	-.31	-.02	
32	..	.09	-.01	-.06	.10	-.02	.06	.40	.17	-.08	-.00	.10	
33	..	-.02	-.03	.16	.17	.40	-.10	.07	-.04	.02	-.06	-.09	
34	..	-.08	.07	.05	.00	.42	.01	.34	.02	.01	-.06	-.08	
35	..	-.14	.05	-.08	-.18	.06	-.38	.09	.06	.10	-.07	-.04	
36	..	.01	.12	.03	.12	.02	.02	.00	-.04	.60	.21	-.10	
No. in hyper-plane $\pm .10$	..	20	17	22	21	24	25	23	30	21	27	24	
No. in hyper-plane $\pm .05$	..	12	13	14	11	19	11	14	21	14	12	13	

TABLE V. RELATIONS OF ROTATED FACTORS

(a) Reference Vectors in Relation to Unrotated Factors

ROTATION MATRIX Λ'

Reference Vectors		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1	..	-.26	.07	.43	-.58	.27	-.02	.29	.20	.20	-.27	-.14	.27
2	..	-.27	-.28	.06	.31	-.24	-.02	-.20	.44	.39	.53	.03	.13
3	..	-.20	.11	.23	.44	.49	-.06	.31	.12	.40	.41	-.04	.10
4	..	.07	.13	.20	.22	-.30	.59	.20	.03	.55	.24	-.05	-.23
5	..	.21	-.30	.11	.05	.48	-.01	.00	-.17	-.30	-.15	.68	.13
6	..	-.06	.05	.22	.22	-.46	-.33	-.37	-.01	-.56	-.31	.16	-.01
7	..	.07	-.07	.17	.17	-.07	-.13	.53	.21	.36	-.54	.36	.24
8	..	-.02	.14	.12	.05	-.32	-.51	.55	-.14	-.33	.29	-.07	.30
9	..	-.05	.06	-.01	.00	-.13	-.04	-.12	-.39	.42	.17	.76	.17
10	..	-.08	-.04	-.10	-.01	.12	.07	-.19	-.65	.43	-.25	-.50	-.07
11	..	.10	.19	.10	.00	.13	-.28	-.38	.71	.14	.04	-.41	.12
12	..						Omitted						

(b) Relations of Rotated Reference Vectors to One Another, $\Lambda'\Lambda$

	1	2	3	4	5	6	7	8	9	10	11
1	..										
2	..	-.17									
3	..	.15	.43								
4	..	-.09	.37	.32							
5	..	-.04	-.30	-.01	-.41						
6	..	-.29	-.11	-.52	-.42	.11					
7	..	.36	-.03	.19	.12	.24	-.04				
8	..	.04	-.03	.12	-.22	-.11	.24	.14			
9	..	-.16	.18	.09	.18	.36	-.04	.23	-.03		
10	..	.03	-.24	-.02	.13	-.29	-.27	-.18	-.30	.01	.
11	..	.21	.37	.17	-.13	-.39	.03	-.10	-.14	-.46	-.17

FACTOR C

(1 in this series)

EMOTIONALLY MATURE STABLE CHARACTER VS. NEUROTIC GENERAL EMOTIONALITY

26	65	Placid vs. Worrying, Anxious.
30	-62	Absence of Neurotic Fatigue vs. Neurotically Fatigued.
13	56	Calm, Phlegmatic vs. Emotional.
14	-52	Not So vs. Hypochondriacal.
8	-49	Unshakable Poise, Tough vs. Easily Upset.
2	46	Emotionally Stable vs. Changeable.

Variables in hyperplane within $\pm .10$ 20

" " " " $\pm .05$ 12

Goodness of simple structure fit = 3.

This factor is unquestionably the same as that obtained in the original multifactorization study² on older men, also in the earlier study of Burt¹ and the later analyses of neurotics by Eysenck.³⁷ Its absence from the parallel study of young men is commented on below. Again the high points of the pattern are, on the negative side, emotionality, instability, and neurotic manifestations.⁴ The

peculiarity of the women's pattern seems to be the high loading of the worrying form of emotionality and the absence of significant loading in persevering, conscientious behaviour on the positive side. In short, high emotionality in men veers toward delinquency, in women toward anxiety neurosis.

FACTOR A

(2 in this series)

CYCLOTHYMIA *VS.* SCHIZOTHYMIA

1	55	Ready to Co-operate	<i>vs.</i>	Obstructive.
12	50	Good Natured, Easygoing	<i>vs.</i>	Spiteful, Grasping, Critical.
7	43	Attentive to People	<i>vs.</i>	Cool, Aloof.
19	-46	Soft-hearted, Kindly	<i>vs.</i>	Hard, Stern.
11	-43	Trustful	<i>vs.</i>	Suspicious.
15	42	Mild, Self-effactive	<i>vs.</i>	Self-willed, Egotistic.

Variables in hyperplane within $\pm \cdot 10$ 17

" " " " $\pm \cdot 05$ 13

Goodness of simple structure fit = 2.2.

Four variables (1, 12, 7, 15) are identical in this and the top seven of the parallel men's factor A, thus clearly identifying this as the first rather than the second of the two schizoid factors, A and H. The variables that are *not* in the factors, but in their hyperplanes, agree in showing that this is the schizothyme factor in which hostility and suspicion, rather than mere withdrawal, characterize the schizoid pole. Both in the factor pattern itself and in the high relative variance of A to H factors, the women resemble the older rather than the younger (parallel) men's data, but the differences are all very slight.

FACTOR G

(3 in this series)

POSITIVE CHARACTER INTEGRATION *VS.* IMMATURE, DEPENDENT CHARACTER

17	57	Persevering, Determined	<i>vs.</i>	Quitting, Fickle.
6	-48	Responsible	<i>vs.</i>	Frivolous.
20	47	Insistently Ordered	<i>vs.</i>	Relaxed, Indolent.
1	45	Ready to Co-operate	<i>vs.</i>	Obstructive.
4	25	Assertive, Self-assured	<i>vs.</i>	Submissive.
2	24	Emotionally Stable	<i>vs.</i>	Changeable.
27	24	Conscientious	<i>vs.</i>	Somewhat Unscrupulous.
29	-24	Practical, Logical	<i>vs.</i>	Sensitively Imaginative.

Variables in hyperplane within $\pm \cdot 10$ 22

" " " " $\pm \cdot 05$ 14

Goodness of simple structure fit = 2.1.

This is clearly factor G, though with a slightly, but probably significantly different emphasis from the parallel men's study. The new emphasis here is on the

'insistently ordered' and 'assertive' aspect and the regard for people. This may hinge on the pattern of character integration training in women. On the other hand these slight differences could be due to the factor being obtained in purer form here than hitherto, for it looks more psychologically self-consistent and is based on more cases and better reliabilities. The earlier patterns seemed somewhat more 'contaminated' with the general emotionality factor, and, though they indicated 'dependence' on the negative side, did not so clearly indicate 'assertion' on the positive side.

FACTOR E'

(4 in this series)

DOMINANCE VS. SUBMISSIVENESS

28	40	Composed vs. Shy, Bashful.
8	—40	Unshakable Poise, Tough vs. Easily Upset.
14	40	Hypochondriacal vs. Not So.
4	33	Assertive, Self-assured vs. Submissive.
30	30	Neurotic Fatigue vs. Absence of Neurotic Fatigue.
3	28	Attention-getting vs. Self-sufficient.
25	25	Unconventional, Eccentric vs. Conventional.
5	22	Depressed, Solemn vs. Cheerful.

Variables in hyperplane within ± 10 21

" " " " ± 05 11

Goodness of simple structure fit = 2.8.

This factor actually shares only two variables—4 and 5—with the dominance factor in men. The justifications for matching it are: (1) No other factor remotely resembling the important dominance factor is found in women. (2) The over-all psychological sense of the pattern is one of social dominance as indicated by poise, composure, self assertion, demand for limelight, and unwillingness to submit to convention. (3) The differences in the pattern from that of men are in accordance with what we should expect, having regard to the social convention which requires more devious expression for dominance in women. Literature, from the *Taming of the Shrew* to the latest popular article on psychology, instances how hard, tough, independent, and boastful behaviour (the variables found in the men's study, but not here) is discouraged as unwomanly, while hypochondria, neurotic fatigue, and perhaps some attention-getting behaviour become alternative routes to power.

The unity of the pattern is perhaps more readily seen when we match with the submissive pole of the men's E factor the variables found here at the negative pole, namely, bashful, easily upset, submissive, self-sufficient, and conventional. The proof of identity in a situation such as this must depend on tying up the two patterns with some common objective or physiological test, e.g., of the androgen-œstrogen ratio (if this should prove to be a masculinity-dominance factor). Meanwhile, for clarity of discussion and reference, we have recognized the descriptive difference by calling this the E' factor.

FACTOR F

(5 in this series)

SURGENCY VS. DESURGENCY

16 —44 Talkative *vs.* Silent, Introspective.
34 42 Gregarious, Sociable *vs.* Self-contained.
9 —40 Energetic, Alert *vs.* Languid, Slow.
33 40 Frank, Expressive *vs.* Secretive, Reserved.
5 —34 Cheerful *vs.* Depressed, Solemn.
7 34 Attentive to People *vs.* Cool, Aloof.
18 —33 Adventurous, Bold *vs.* Cautious, Retiring, Timid.

Variables in hyperplane within	± 10	24
" " " "	± 05	19
Goodness of hyperplane fit = 7.3.		

Four variables (16, 9, 5, 18) identify this with the F factor in the parallel group of young men. The extra emphasis here on sociability versus aloofness, in variables 7 and 34, is shared with the factor pattern in the older men,² and the factor here has a somewhat lower variance as in the older men. Otherwise the pattern has the stability and clarity typical of the surgency-desurgency source trait. A misogynist might note that 16—Talkativeness—has the greatest divergence of position in the two patterns, rising from fifth in the men to first place in the women, the other common variables retaining their order.

FACTOR K

(6 in this series)

TRAINED, SOCIALIZED, CULTURED MIND VS. BOORISHNESS

10 —56 Intellectual, Cultured vs. Boorish.
21 57 Polished vs. Clumsy, Awkward.
31 41 Æsthetically Fastidious vs. Lacking Artistic Feeling.
35 —38 Independent-minded vs. Dependent, Immature.
29 32 Sensitive Imaginative vs. Practical, Logical.
9 —25 Energetic, Alert vs. Languid, Slow.

Variables in hyperplane within	$\pm \cdot 10$	25
" " " "	$\pm \cdot 05$	11
Goodness of simple structure fit = 2·5.		

Three of the six highest in this pattern are the same as those of the six highest (10, 21, 29) in the factor in the parallel men's study. All are consonant with the main interpretation of K and with the pattern in the original, older men's study,² which also had æsthetic interests and "smart, independent mind" (absent, however, from the younger men's group). The only difference worthy of comment, therefore, is the greater variance this factor has than in the men's studies.

FACTOR H

(7 in this series)

ADVENTUROUS CYCLOTHYMIA *VS.* WITHDRAWN SCHIZOTHYMIA

32	40	Marked Interest in Opposite Sex	<i>vs.</i>	Slight Interest in Opposite Sex.
31	34	Æsthetically Fastidious	<i>vs.</i>	Lacking Artistic Feeling.
34	34	Gregarious, Sociable	<i>vs.</i>	Self-contained.
28	30	Composed	<i>vs.</i>	Shy, Bashful.
21	24	Polished	<i>vs.</i>	Clumsy, Awkward.
7	23	Attentive to People	<i>vs.</i>	Cool, Aloof.

Variables in hyperplane within ± 10 23

" " " " ± 05 14

Goodness of simple structure fit = 1.6.

Three of the four highest variables here are the same as three of the four highest in the parallel men's study.⁵ The intruding variable here—æsthetically fastidious—had positive loading in some alternative rotations of the earlier study⁴ and, unless we are dealing merely with a sex difference in this respect, the present pattern is more to be accepted, having a better hyperplane and being based on a larger sample, etc. Both this and the parallel study differ from the older adults² in having less emphasis on the kindly and idealistic behaviour and more on interest in the opposite sex, as if representing the same pattern prior to sublimation. Otherwise the pattern is essentially the same: a simple, withdrawn schizoid quality on the negative pole and an adventurous sociability and sexuality at the positive pole.

FACTOR N

(8 in this series)

GENTEEL SOPHISTICATION *VS.* ROUGH SIMPLICITY

21	78	Polished	<i>vs.</i>	Clumsy, Awkward.
7	—57	Cool, Aloof	<i>vs.</i>	Attentive to People.
31	42	Æsthetically Fastidious	<i>vs.</i>	Lacking Artistic Feeling.

Variables in hyperplane within ± 10 30

" " " " ± 05 21

Goodness of simple structure fit = 2.3.

This factor pattern has no immediate resemblance to any known factor. Two of the items—7 and 21—are those which, owing to defective estimates of communality, alone exceeded unity in the finally obtained communality. The factor is also peculiar in having no loadings greater than chance for anything but these three variables, though in these the loadings are conspicuous. Until further research shows whether this is an artifact due to the above causes, no serious effort should be made to interpret it; but it is certainly possible to recognize individuals with a high endowment in a cluster of the kind, and one recognizes it as the apparent product of genteel and sophisticated upbringing, in a specialized social class, quite distinct in meaning from the K factor, which is apparently general education.

FACTOR B

(9 in this series)

INTELLIGENCE (IN TOTAL PERSONALITY) *VS.* MENTAL DEFECT

36	60	Intelligent <i>vs.</i> Unintelligent.
17	26	Persevering, Determined <i>vs.</i> Quitting, Fickle.
8	-23	Unshakable Poise, Tough <i>vs.</i> Easily Upset.

Variables in hyperplane within	$\pm \cdot 10$	21
" " " "	$\pm \cdot 05$	14
Goodness of simple structure fit = 1.3.		

This factor seems 'purer' in the sense of having a higher loading in the intelligence test and lower loading in any other variables than the two previous B factors. Its hyperplane and goodness of fit, however, are not better, so that conceivably the difference is partly due to approximations in rotation. What variables have any loading are of the usual 'characterial' kind. One must conclude, however, that there are indications that in this group, either because they are women or because they tend to come from a somewhat higher economic group than the parallel group of men, qualities of conscientiousness and perseverance are less involved in the intelligence factor.

FACTOR M

(10 in this series)

SPIESSBURGER CONCERNEDNESS *VS.* BOHEMIANISM

25	-37	Conventional <i>vs.</i> Unconventional, Eccentric.
31	-31	Lacking Artistic Feeling <i>vs.</i> Æsthetically Fastidious.
29	-27	Practical, Logical <i>vs.</i> Sensitively Imaginative.
27	22	Conscientious <i>vs.</i> Somewhat Unscrupulous.
36	21	Intelligent <i>vs.</i> Unintelligent.
8	-20	Unshakable Poise, Tough <i>vs.</i> Easily Upset.
26	-17	Worrying, Anxious <i>vs.</i> Placid (not Anxious).

Variables in hyperplane within	$\pm \cdot 10$	27
" " " "	$\pm \cdot 05$	12
Goodness of simple structure fit = .6.		

This has three variables the same as those of the highest six in the parallel men's study⁵ factor M, while the rest are psychologically consistent. Thus both artistic interests and lack of conscientiousness were discussed in connexion with the previous M factor, though they definitely appear only in this more exact study with larger population.

FACTOR D

(11 in this series)

INFANTILE, STHENIC EMOTIONALITY *VS.* PHLEGMATIC
FRUSTRATION TOLERANCE

24	30	Demanding, Impatient	<i>vs.</i>	Emotionally Mature.
3	28	Attention-getting	<i>vs.</i>	Self-sufficient.
22	27	Prone to Jealousy	<i>vs.</i>	Not So.
21	—26	Clumsy, Awkward	<i>vs.</i>	Polished.
15	—23	Self-willed, Egotistical	<i>vs.</i>	Mild, Self-effacing.
27	—21	Somewhat Unscrupulous	<i>vs.</i>	Conscientious.

Variables in hyperplane within $\pm \cdot 10$ 24" " " " " $\pm \cdot 05$ 13

Goodness of simple structure fit = 1.2.

This factor has recognizably the same pattern as D in the older men's study, variable 21 being the only novelty and "neurotic symptoms" the only absentee variable. But it also shows two variables, 24 and 3 appearing in the I factor in the parallel group of younger men. In the latter study⁵ it was explicitly noted that I had some resemblance to D, a resemblance again evident here. We thus have one study (young men) which yields I and not D, another (young women) which yields D and not I, and a third (older men) yielding D and I simultaneously. The most probable conclusion is that D and I are what have been called "co-operative factors,"⁴ partly similar in loading pattern and not readily separable in most populations. In the present research this factor has the smallest variance of all factors and we cannot hope, in the circumstances, to separate I out from it. The I factor shares the first two variables above, as stated, but then runs, on the positive side, into 'frivolous,' 'aesthetically fastidious,' 'sensitively imaginative,' 'soft-hearted,' 'dependent,' and 'hypochondriacal.' By contrast the present factor, D, seems to centre upon the sthenic emotionality described by Burt.¹

The rotation of the twelfth factor could not be satisfactorily terminated and it was decided to leave it on indeterminate residual, with 24 variables in the hyperplane.

VI. SUMMARY

(1) A factorization of 36 traits, representative of the total personality sphere, for 240 college women averaging 20.7 years of age has yielded eleven or twelve general factors, the twelfth being doubtful and practically a residual.

(2) Although the search for simple structure was prolonged and difficult the simple structure eventually attained was very satisfactory, more than 20 out of 36 variables being in the factor hyperplane (within $\pm \cdot 10$ correlation) for all but one factor, and the same structure being lost and regained during the 588 rotations.

(3) Ten of the eleven factors with adequate variance, namely, A, B, C, D, E, F, G, H, K, M, are unmistakably the same as factors found previously in two groups of men. (All but C, D, and M appeared in *both* previous studies.) In all cases

identification rested on three or more of the six or seven highest variables being the same in the two factors matched. The probability of three given variables out of the total of 36 appearing in any group of six by chance is well below the 1 per cent. level. ($^{33}C_3$ divided by $^{36}C_6$ equals 1/357.)

(4) Unmatched factors consist of I, J, and L present in men and absent in women, and N present in women and absent in men. Factors I and J may have failed of separation here because of too small variance, for there are suggestive traces of them. Factor N, 'Genteel sophistication,' seems likely to be a product of women's education patterns. Factor L, 'Paranoia,' may be absent here for the same reason that makes the clinical paranoia syndrome rarer in women than men.

(5) Some differences of loading emphasis exist between the patterns of factors found for men and women, even when the factors are overwhelmingly similar and presumably generically identical; but it is not possible yet to say how much of the observed difference is systematic and how much due to chance. However, these differences, commented upon while describing the individual factors above, make good sense psychologically and may well be real and systematic. In addition to what is thus explicable in terms of the present psychological understanding of masculinity-femininity, there is a general and not so readily explicable tendency for the women's factor patterns to resemble those of the older men (average age 30 years) rather than men of the same age (20 years).

(6) The differences of loading emphasis are greatest in C and E and least in A, H, and K. In C, General Emotionality (obverse), the women's factor has more emphasis on fearful emotionality and neurotic traits. In E, Dominance, the men's factor lacks emphasis on attention-getting and hypochondria and loads more powerfully toughness and boastful self-assertion.

(7) With the possible exception of factor E, all of the principal dimensions of personality could be assessed on behaviour, rating variables common to the definition of these dimensions in men and women. Only a few factors of small variance seem entirely peculiar to men or women.

The writer wishes to express his thanks to Eleanor Maxwell, Betty Pollard, and Carole Romer for careful statistical work and perseverance with the rotation problem.

REFERENCES

1. Burt, C. L. : *The factors of the mind*. London : Univ. of London Press, 1940.
2. Cattell, R. B. : "The description of personality : principles and findings in a factor analysis." *Amer. J. Psychol.*, LVIII, 1945, 69-90.
3. Cattell, R. B. : 'The diagnosis and classification of neurotic states : A re-interpretation of Eysenck's factors.' *J. Nerv. Ment. Dis.*, CII, 1945, 576-589.
4. Cattell, R. B. : *The description and measurement of personality*. New York : World Book Company, 1946.
5. Cattell, R. B. : 'Confirmation and clarification of the primary personality factors.' *Psychometrika*, 1947.
6. Cattell, R. B., and Saunders, D. R. : 'The relation of objective test, questionnaire and behaviour rating personality factors in men.' (To be published.)
7. Eysenck, H. J. : *Dimensions of personality*. London : Kegan Paul, 1947.
8. Guilford, J. P., and Guilford, R. B. : 'Personality factors, D, R, T, and A.' *J. Abnorm. Soc. Psychol.* XXXIV, 1939, 1.
9. Symonds, P. M. : *Diagnosing personality and conduct*. New York : Appleton Century, 1931.
10. Thurstone, L. L. : *Primary mental abilities*. Chicago : Univ. of Chicago Press, 1938.
11. Thurstone, L. L. : *Multiple factor analysis*. Chicago : Univ. of Chicago Press, 1947.

BOOK REVIEWS

Dimensions of Personality. By H. J. Eysenck, Ph.D. (London). Kegan Paul, Trench, Trubner & Co. Ltd. Pp. xii + 308. 25s. net.

This book is described as "the result of a co-operative effort to discover the main dimensions of personality and to define them operationally by means of strict quantitative procedures." The general aim has been to test the possibilities of applying to normal and subnormal adults the various experimental and statistical procedures that have proved so valuable in the study of normal and subnormal children. Psychiatrists have long been urged by psychological writers to check their methods and conclusions by adopting modern statistical devices; and here is a book which will show that the suggestion is actually practicable. A special feature of the work has been the close collaboration between non-medical psychologists and medically qualified psychiatrists.

The majority of the persons tested were patients drawn from the Services. It is emphasized that the term 'neurotic' as applied to such groups does not possess the same meaning of civilian patients. "The 'neurosis' exhibited was not so much an illness as a simple failure to adapt to Army routine and discipline." This naturally limits the generalizations that can be drawn.

The investigators "started from the position that measurement in the field of personality is impossible until the dimensions along which such measurement can take place are known." In accordance with this principle they determined to carry out "a large scale factorial study on a variety of personality traits whose presence or absence in 700 'neurotic' cases could be recorded by the psychiatrist in charge"; and their "general view of factor analysis" has, we are told, determined both the general plan of their enquiry and the way in which their data are reported and discussed. "Our own position," they say, "is very similar to that of Burt": with him they regard factors as "principles of classification described by selective operators" rather than as necessarily identifiable with "primary abilities," "fundamental traits," "temperamental types," and the like.

Most of their analyses have accordingly been carried out either by the method of simple summation or by the method of group factor analysis, leading in either case to a factor matrix which includes a general factor as well as bipolar or group factors. Indeed, they have re-analysed data factorized by American writers, such as Guilford, Cattell, and others, to show that a scheme which includes a general factor is, as a rule, more plausible in dealing with personality-assessments than one which, like 'simple structure,' tends to exclude or minimize the general factor. A suggestive appendix reporting 'an experimental study in the methodology of factor analysis' seeks to establish this conclusion more firmly. The final upshot of their work is that "a hierarchical view of personality structure, akin both to Burt's conception of factors and McDougall's theory of sentiments, would appear to account for the known facts better than any other." A thorough analysis of the data leads, so they claim, to "the discovery of two main factors," which, "subject to various reservations, may be labelled 'neuroticism' and 'extraversion-introversion.'"

Their supplementary researches have been concerned with ascertaining the best experimental procedures for testing and assessing the 'dimensions' thus discovered. A large number of procedures—questionnaires, rating-methods, tests of temperament, ability, suggestion, and appreciation, and measurements of body type—have been employed; and their discriminative value examined in detail. Some readers may feel a little disappointed that the investigators were not also able to include in their scheme of enquiry the method of "observation in standardized situations"—a method which with children has given reliable results, and which, according to those who have adopted it in personnel selection

for the War Office and Civil Service, is said to be at once more reliable and more valid with adults than any set of personality-tests used by itself.

The sections of the book which are most likely to arouse questions in the minds of the readers experienced in this field are those discussing the diagnostic value of the methods proposed. Several critics have pointed out that, in their earlier publications, the investigators seemed a little too ready to infer that, if the mean scores obtained from the normal and neurotic groups with any given test were statistically significant, then that test might be used to discriminate between the two groups. In his discussion of the significance of body build and similar traits, Dr. Eysenck endeavours to forestall such criticisms by emphasizing the fallible nature of the data available. "In answer," he says, "we may point to the comparative lack of reliability of the psychiatrist's ratings"; and he argues that, "if in spite of this lack of reliability, a significant difference is found, then we may consider the actual difference to be greater."

In their last chapter, giving "A Synthesis and Conclusions," the writers tabulate "the main differences found between persons at the opposite ends of the respective dimensions," i.e., of the scale running from normality to neuroticism and from introversion to extraversion. "A comparatively large number of tests were found to be discriminative in these connexions;" and the nature of these, and of the non-discriminative tests, is made the basis of "heuristic hypotheses" regarding the nature of the statistical factors discovered.

To assess discriminative value, two different formulæ were first employed. "In the first place," we are told, "an Index of Screening Efficiency was elaborated": this seems to be simply what Yule calls the 'product-moment correlation for a 2×2 -fold table,' which assumes the frequencies to be concentrated at single points (a somewhat doubtful assumption with traits whose distribution is admittedly continuous and presumably unimodal). The second formula was Hunt's 'Selection Index.' By these or other criteria a battery of 10 tests was finally selected and administered to 105 neurotic patients and 93 surgical cases. The correlation between the total scores, reduced to an 8-point scale, and the psychiatric classification into neurotic and non-neurotic, proved to be 0.73. This figure is far higher than those usually reported with tests of the type employed; and it would therefore be a little more satisfying to know how the correlation was calculated: if it is a biserial correlation, were the requisite conditions fulfilled?

A further method of assessing the correlation was also adopted. The correlation between two sub-batteries, consisting of 5 tests each, was calculated, and found to be 0.60. This is "corrected by the Spearman-Brown prophecy formula"; and thus raised to 0.75. From this it is inferred that the correlation of the test battery with a perfect criterion would be $\sqrt{0.75} = 0.87$. This is modestly described as "a comparatively high value" for such tests; and it is suggested that the actual validity of the tests probably lies midway between the two figures thus reached. The final conclusion is that "tests of temperament and personality may have reached a stage where they can take their place with the usual tests of cognitive function and special abilities."

This conclusion, like the tone of the whole volume, is perhaps a little sanguine. If, therefore, any practical psychologist is thinking of using such tests for purposes of personnel selection or psychiatric diagnosis, he would be wiser to wait for a repetition of the experiments by wholly independent investigators. Time after time we have heard of equally promising results claimed for similar test-procedures only to be sadly disillusioned when the first hopeful enquiry was repeated. This is merely a warning to the practical worker. The theoretical psychologist will form his own conclusions; and will assuredly find much in this book that is at once stimulating and suggestive. The industry and enthusiasm of the whole team, and the success with which they have organized a co-operative scheme of work, deserve unhesitating praise.

E. K.

Fundamentals of Statistics. By T. L. Kelley. Harvard University Press. Pp. xvi + 756. 47s. 6d. net.

Kelley's textbook of *Statistical Method*, when it first appeared in 1923, gave what had long been wanted—a clear, comprehensive, and systematic account of the methods developed by what Kelley himself termed “the English School,” that is, by “the master-analyst, Karl Pearson, and his co-workers.” For illustrative “problems in need of statistical analysis” Kelley went to Edward L. Thorndike; and thus his book was particularly well adapted to the needs of those undertaking statistical research in the field of psychology and education. His new volume contains all the admirable features of the old; but, as he points out, it is something more than a mere revision of the earlier work. While so many books on statistics have unfortunately made little or no attempt to give the reader any insight into, or explanation of, the methods and formulæ they employ, the present volume lays a welcome stress on the logical principles underlying statistical arguments.

In the earlier chapters, which are intended to be elementary, such topics as the presentation and tabulation of data and the use of graphical methods are discussed in a clear and instructive fashion. Later chapters deal with measures of variation and central tendency, the properties of the normal distribution, the statistics of attributes, Karl Pearson's chi-squared technique, correlation and regression problems (bivariate and multivariate), and various other subjects. Among the latter it is gratifying to find a brief discussion of sequential analysis. At the end of the book there are helpful notes on various mathematical techniques employed in the theory of statistics—such as determinants and matrices, the solution of equations, the use of Lagrange multipliers, of growth curves, and the like. Finally, there is a short collection of statistical tables.

It is perhaps too much to expect of any book on statistics by a single author that it should do equal justice to all parts of the subject; and it is inevitable that there should be a discernible bias in certain directions. The present volume will probably prove particularly useful to workers in the field of education or psychology, where multivariate analysis plays a large and important part. Nevertheless, it seems a pity that more attention has not been given, for example, to the analysis of variance, a method whose value is becoming increasingly well realized.

Factor analysis is formally defined in the useful appendix of definitions; and in the text we are told that the relations connected with this subject can usefully be expressed in matrix notation. But it is surprising to find that these brief allusions are all that the volume contains on a subject to which Kelley himself has made notable contributions. The reviewer would also have liked to see more discussion given to the theory of probability and its relevance to tests of significance. Since this is a book on the foundation of statistics, some consideration might also have been given to the theory of estimation beyond the references to R. A. Fisher's work in this field.

The author is careful to distinguish between population parameters and their estimates from a sample. The notation adopted appears to the reviewer, however, to be somewhat cumbersome and inconvenient. The plan of reserving, so far as possible, Greek letters for population parameters and Roman letters for sample estimates has now gained fairly wide acceptance, and has been found satisfactory in practice.

In spite of these minor criticisms, the book should prove extremely helpful both to beginners and to research students. It contains many suggestive explanations of difficult questions. The treatment throughout is at once lucid and logical; and the selection of novel topics original and instructive.

D. N. L.

Description and Measurement of Personality. By R. B. Cattell. New York, World Book Company. Pp. xx+602. 15s.

This book is described as the first of two volumes, which together might be entitled 'The Empirical Study of Personality.' The phrase, we are told, was "intended to express, quite militantly, a definite break with the majority of writings having intuitive foundations, which hitherto have commanded the field in the clinical study of the subject." Volume II will be 'developmental in approach'; the present volume is essentially 'cross-sectional.' It has a twofold object—"to bring into perspective the knowledge so far obtained, and to suggest problems and methods for further research." It is thus at once a survey of methods and of results.

In his endeavour to present a synoptic review of existing results, Dr. Cattell has carried out an amazingly thorough search of the available literature; and has sought to catalogue, in the form of a detailed summary, all the principal 'surface traits' and 'source traits' that have been discovered by the various methods at present in vogue, namely, 'behaviour ratings,' 'self-inventories,' and 'objective tests.' From this catalogue he eventually deduces a short series of twelve 'primary' source traits, which he considers to be now well established. For admission to this final list, a factor must have been found in at least three independent researches.

In his scheme of the general structure of personality, Cattell appears to agree with most workers in this country. He distinguishes (a) innate or constitutional traits from (b) those that are acquired, and recognizes three main aspects, namely, (i) cognitive abilities, (ii) temperamental characteristics, and (iii) conative or motivational tendencies. His twelve primary traits include so-called 'general factors,' such as 'general intelligence' and 'general emotionality,' and a number of temperamental, emotional, or psychoneurotic factors of narrower range, such as 'cyclothymia,' 'surgency,' 'dominance,' 'anxiety,' 'neurasthenia,' and their opposites. The descriptive labels given to each indicate an attempt to envisage them as essentially bipolar in form, and suggest that much of the material has been gathered with the psychoneurotic rather than the normal personality in mind.

The personality traits, however, do not include 'special aptitudes or abilities.' Of these, Cattell believes that twelve are 'more basic'; but he thinks another half dozen or so should be recognized as "established by present research." Both the scheme and the list which he gives in his book coincide (with one or two exceptions—e.g., the omission of imagery) fairly closely with those of Burt in *The Measurement of Mental Capacities* (1926). His whole review of the factual data, supplemented as it is by a bibliography of nearly 300 references, constitutes a mine of information, which will be invaluable to future investigators.

The statistical psychologist will doubtless be still more interested by Dr. Cattell's discussion of methods than by his provisional summary of results. Here he shows himself to be far more eclectic and impartial than most writers on this subject. Further, he is one of the few psychologists who seem equally well acquainted alike with British and with American work. Instead of appearing as the champion of any one school, he takes what is most plausible from each of them in turn, adds much that is both original and suggestive, and endeavours (with the aid of a somewhat formidable terminology of his own) to combine them into a "coherent but critical synopsis."

His main suggestions in regard to mathematical procedure are brought together towards the end of his book in the shape of a "systematic presentation of theorems." He begins with what he terms the 'basic theorem,' namely, that "every behaviour manifestation of personality, PR (Personality Response), for an individual i , can be predicted by the following type of equation :

$$PR_{j,i} = S_{1,j}T_{1,i} + S_{2,j}T_{2,i} + S_{3,j}T_{3,i} \dots + S_{n,j}T_{n,i} \dots T_{j,i} + e_j \quad (i)$$

where $S_{1,j}$, $S_{2,j}$, etc., are factor loadings, and T_1 , T_2 , etc., are source traits." Of the T 's, some are "general to many performances," others apparently limited to groups of performances; and at least one, T , is "peculiar to the situation in which PR_j is measured"; while e is "some experimental error of measurement, not function fluctuation." It will be observed that, unlike Thurstone, he has no hesitation over accepting 'general' factors;

and, unlike Spearman, he is ready to recognize a large number of 'group' or 'bipolar' factors. But he rightly insists that "the division between general and group factors is relative rather than absolute," and that, "except in abilities, we should not expect either general or group factors to be wholly positive." Consequently, he does not reject bipolar factors involving negative saturations. On the other hand, he fully agrees with Thurstone that "personality source traits seem themselves to be not independent, but correlated."

Now I venture to suggest that for the investigator the basic problem is not so much to "predict" the "personality responses," as to predict, or rather to estimate, the "source traits." This Dr. Cattell discusses more briefly. He describes two modes of calculation—a theoretical or 'precise' procedure, and a rough and ready practical procedure.

For theoretical purposes he gives a regression equation of the usual type, namely,

$$T_{pi} = \beta_{pa} \cdot PR_{a,i} + \beta_{pb} \cdot PR_{b,i} + \dots + \beta_{pn} \cdot PR_{n,i} \quad (ii)$$

where T_{pi} represents "the personal endowment of individual i " in regard to the p th "source trait," and $PR_{a,i}$, etc., are described as the individual's "endowments in each of the performances": (unless I have misunderstood him, however, 'endowment' as applied to PR must be a slip for 'achievement' or 'measurement').

Then, turning to the more practical methods of assessment, he observes that "in applied psychology, such estimation by weighted scores usually involves more trouble than is justified by the slight gain in accuracy. It is sufficient to find half a dozen or more variables heavily loaded in the factors concerned, and little in any other, or with mutual suppressors in others: these are weighted with unity, and all others zero" (except, of course, the 'suppressors,' which will presumably have weights of minus unity).

His discussion of these principles, and the corollaries that he draws from them, have provoked considerable criticism both from a theoretical and from a practical standpoint. The theoretical critic objects that many of the inferences deduced hold good only if the factors or 'source traits' are assumed to be orthogonal, which Cattell has explicitly denied. Certainly, his somewhat condensed account of 'factor patterns,' 'factor structure,' and the requisite 'loadings' (pp. 559-560) is exceedingly difficult to accept as it stands. But Cattell himself warns us that here he is writing for "readers who will not be primarily statisticians"; and it may be that some of the apparent confusion is due to a desire to present a technical argument in a non-technical form. His suggestions for practical assessments have invited criticism from another quarter. If after all it is sufficient just to "average the group," why the preliminary mathematical calculation, and why the elaborate theorems? If, on the other hand, the preliminary mathematical analysis is essential to scientific rigour, why fall back, at the end of it all, upon simply totalling or averaging the measurements just as they stand?

Cattell's reply apparently is that the preliminary factorial calculations are needed to show which traits deserve selection for the final process of averaging: "the mathematical work reminds us that the choice is not arbitrary, but directed by factor analytic findings." But he might, I fancy, carry his defence still further. His answer appears to concede that the use of simplified or even arbitrary weights, and the acceptance of unstandardized marks, necessarily entail a sacrifice of mathematical rigour. But this, I would suggest, need be by no means true.

Such a view perhaps demands a brief demonstration. We may fully agree that the 'basic theorem' should deal with the factorization of the measurement matrix rather than of the correlation matrix. Moreover, if we re-write his equation (i) (adopting matrix notation with the usual symbols) in the form

$$PR = M_0 = M_1 + M_2 + \dots + M_r \quad (ia)$$

where $M_i = \frac{1}{v_i} f_i p_i$ (his $S_i T_i$),

then (as I have shown elsewhere) we may take

$$p_1 = w'_{1a} M_0, f_1 = M_0 w_{1b}, \text{ and } v_1 = w'_{1a} M_0 w_{1b} \quad (iii)$$

where the w_{1a} and w_{1b} may each be *any* column of weights or 'loadings' (provided, of course, they do not reduce the divisor, v_1 , to zero). M_1 will thus be 'hierarchical' (of rank one). And, on applying the same procedure to each of the residual matrices in turn, we shall obtain r linearly independent factors. According to our aims we may seek to maximize the w 's; or we may treat them as consisting of +1 only; of +1 or -1; or of +1 and zero; and so on. Again, we may prepare M for the weighting process by subtracting means, by normalizing the deviations, or by first reducing M to a square symmetrical form by working with $R = MM'$. The omission of these familiar restrictions in the interests of practical simplification may or may not impair the theoretical interpretation or the scientific accuracy of the measurements as finally computed. But it is clear, from the freedom allowed us in choosing our w 's, that it need not destroy the mathematical rigour of the analysis.

In identifying factors, and in rotating the initial factors to obtain a consistent structure, he considers the most fundamental principle to be that of 'proportional profiles.' Factors are the same when their saturations are proportional. We are therefore to rotate axes until this proportionality is secured. But, when several factors are involved, this requirement, he believes, will be difficult to secure without "harassingly numerous" equations. Here I venture to suggest that the 'symmetry criterion' (which was originally deduced from the proportionality principle¹) might serve his purpose.

Although at times his work shows signs of haste and has apparently been somewhat hurriedly revised, it is unquestionably as full of ideas as it is of data. It forms a valuable contribution to its subject; and can be neglected by no writer or investigator who desires to enter the field of individual psychology.

C. B.

¹ Burt, C. 'The Unit Hierarchy and its Properties,' *Psychometrika*, III (1938), p. 161 and refs.

AN ITEM-ANALYSIS OF THE TERMAN-MERRILL REVISION OF THE BINET TESTS

By RAYMONDE COLE

Department of Psychology, University College, London

I. *Data.* II. *Procedure.* III. *Results.* IV. *Summary and Conclusions.*

I. DATA

The Adaptation of the Revised Scale for English Children. Shortly after the publication of the Terman-Merrill Revision of the Binet tests (6) a committee was set up to adapt and standardize the new revision for use in this country (7, p. 255). Representatives of practically all the child guidance centres at that time established, and nearly all departments interested in educational psychology, took part in the discussions; and plans were drawn up for investigating what modifications were needed in the instructions and the wording,¹ and for collecting data to check the age-assignments and re-allocate the tests where necessary. With the co-operation of teachers, educational psychologists, and research-students, a large amount of information has been compiled. But its analysis has been interrupted by the war.

In the main the investigation has been confined to the first of the two alternative scales (Form L). For this sufficient material has been accumulated to determine the approximate order of difficulty of the several tests for English children; and the new age-assignments (based on the 50 per cent. criterion, 2, p. 152) have now been published (12). In a discussion of the points calling more particularly for investigation (7, p. 256), Burt has drawn attention to the wide differences in validity shown by the individual test-problems included in the earlier versions, and has argued that "an intensive item-analysis for each test in the new scale is an urgent requisite." A detailed study of this sort would seem essential before the scale is adopted for use in this country, since (as numerous researches have shown) items which have appeared highly satisfactory in France or America may have little or no diagnostic value with British children.

The Number of Cases. The total number of individuals for whom detailed test-results are now available amounts to over two thousand. This includes between 150 and 200 boys or girls for every age from 7 to 14, about 100 young people from every age from 14 to 21, and about 50 to 100 children in the pre-school age-groups. For most of the children of school age, teachers' assessments of intelligence, usually on a sixfold scale, were also secured.

¹ Two English revisions have been made by the Committee. The first was a version drawn up partly by correspondence and partly by round-table discussion, based on general experience. This was circulated for use in roneo'd form; and on the basis of further trials and criticisms a second revision has been made. This part of the work has been carried out mainly by Sir Cyril Burt with the assistance of Miss Constance Simmins (Secretary of the Committee) and later Miss E. L. Matthews, Miss M. Davidson, Miss E. John, Miss E. Denton, and others. With the kind permission of Professor Terman and his publishers, arrangements have been made to publish this revision as soon as it is complete; but unfortunately the material for certain of the tests (notably the Absurdities and Memory for Sentences) has been found still to need further modification. In preparing this paper I have made full use of an earlier memorandum drawn up in 1939 for the Committee by Miss Matthews, giving a preliminary item-analysis for ages 8 to 13.

II. PROCEDURE

Variety of Methods. The notion of item-analysis as an adjunct to test-construction appears to have been first put forward by Burt, and was first carried out in detail in his L.C.C. Report (1921) on the original Binet-Simon tests and on group-tests of general intelligence (1, 2). The novel feature of the method was the systematic investigation, not merely of the value of any new test taken as a whole, but also of the suitability of each separate item by both quantitative and qualitative methods: the quantitative study consisted essentially in calculating the correlation between the item and various independent criteria of intelligence; but this was supplemented by a study of its difficulty, ease of marking, lack of ambiguity in the question and in the replies, and variations with sex, social status, etc. (see 2, Appendices II and III, pp. 230f., 236f.). The phrase 'item-analysis' was originally introduced to cover the whole of this intensive study; but has more recently come into vogue as a convenient technical term to denote the employment of statistical methods for studying item-validity only. It will be used in this narrower sense throughout the present article.

In America the proposal was taken up in Vincent's monograph on 'The Study of Intelligence Test Elements' (*Teachers' College Contributions to Education*, No. 124, Columbia University, 1924). And since that date numerous publications on the subject have appeared, and numerous alternative procedures have been put forward.¹ The method to be adopted here, therefore, requires some preliminary defence. In a systematic review of the subject (11) Burt concludes that "there can be no *one* best procedure; what is best in any particular case can only be decided by a careful consideration of (i) the special aim in view and (ii) the type of data available: this implies that we must first be clear as to the particular kind of enquiry for which this or that formula is really appropriate—a point seldom discussed in the summaries of the various formulæ suggested."

Conditions Influencing Choice of Procedure. The full product-moment formula will rarely be applicable. The assessments obtained from the separate test-items are nearly always in 'dichotomous' form: i.e., the responses are divided into two categories—'right or wrong,' 'pass or fail,' etc. We have first of all, therefore, to decide whether this dichotomous classification is to be regarded as derived from (a) a continuous *normal* distribution or (b) from a discontinuous *point* distribution. In general, it will be natural to assume that the distribution underlying the item-responses is really the same as that underlying the criterion.

The assessments which provide the criterion may take the form of either (A) *graded* measurements, or (B) a *dichotomous* classification. Graded measurements, though at times actually exhibiting an asymmetrical distribution or (as in the use of ranks) a rectilinear distribution, are commonly assumed to be samples from, or convertible into, a normal distribution. In the absence of any definite reason to the contrary, the dichotomous classification can be considered as obtained by dividing a continuous normal distribution into two sections: indeed, often it is actually derived in that way. In certain cases, however, it is better treated as consisting of two point-distributions (for example, when the two contrasted groups have been drawn from the extreme ends of a normal distribution, and the middle section has been omitted).

It needs no statistical knowledge to see that, whether the criterion be graded or grouped, a plausible measure of efficiency can be based on the quantitative *differences* between groups. When the criterion consists of graded marks, we shall naturally contrast the average mark of those who pass with the average of those who fail. When the criterion merely divides the tests into two groups, the bright and the

¹ See 4, pp. 428–437; also Adkins, D. C., *Psych. Bull.* XXXV, 1938, pp. 655f. A recent bibliography of the subject runs to over 100 numbers (11).

dull, for example, we shall naturally contrast the percentage who pass among the bright with the percentage who pass among the dull.¹

Items like those in the Binet scale, however, vary widely in difficulty. Owing to these differences in difficulty, even those which are suitable for the same age-group will divide the group at different levels and therefore in different proportions. And when the same item is applied to different age-groups, the proportions will differ owing to the differing ability of the groups. Further, when we are concerned to compare results from different groups, the marks or measurements that form the criterion may fluctuate considerably in variability (as estimated by variance or standard deviation) from one group to another. Hence in these cases it will be desirable to eliminate, or at least allow for, (i) the effects of differences in difficulty (i.e., in p or q , to adopt the notation used below) and (ii) the effects of differences in variability (i.e., in σ). This entails the use of what have been called *correlational* formulæ.

These considerations lead to the following scheme of procedures. The classification is abridged from that suggested by Burt (11).²

A. CRITERION GRADED (TWO-ROW FREQUENCY TABLES).

1. Correlational Measures.

(a) Item Assumed to follow a Normal Distribution.

(i) *The Normal Biserial Correlation* (Pearson's formula³ : $r_{bis} = \frac{d}{\sigma} \cdot \frac{pq}{z}$). This formula assumes that the dichotomy furnished by each item rests on a normal distribution. On theoretical grounds, it is claimed (11), "this would be the best procedure provided the criterion-distribution does not deviate far from normality. In practice, however, particularly where comparatively small groups have been tested, the deviations may be appreciable; and the formula is then apt to give erratic or exaggerated figures, sometimes exceeding 1.00."

(b) Item Assumed to follow a Point-Distribution.

(ii) *The Point Biserial Correlation* ($r_p = \frac{d}{\sigma} \sqrt{pq}$). This is virtually a 'discriminant function for a single item' (cf. 10, p. 225). It assumes that the dichotomy furnished by the item-responses rests on a point-distribution, although the criterion does not—a somewhat exceptional type of case.

2. Difference Measures.

(iii) *The Standardized Difference between Means* ($\frac{d}{\sigma}$). This was suggested by

¹ These simple devices were indeed frequently used by Binet and others in the early days before British psychologists had familiarized investigators with Pearson's various methods for measuring correlation. For the defects of these earlier expedients and the effects of disregarding difficulty, etc., cf., for example, the instances cited by Burt (2, p. 219).

² The formulæ given here are not necessarily the best for computing purposes. Several more convenient versions are given in 11.

³ It does not seem to be generally realized that this method of measuring correlation (like most others) was devised by Pearson (*Biometrika*, VII, 1909, p. 97). A paper has even been published discussing the merits of 'Biserial r versus Pearson r as measures of item-validity' (*J. Educ. Psych.*, XXVII, 1930, pp. 471-2), thus suggesting that Pearson was responsible for the latter but not the former.

Burt as an alternative method of expressing 'percentage overlapping'¹; and is especially useful when p is approximately constant, or where the test-item is assumed to yield two point-distributions.

(iv) *The Simple Difference between Means* ($d = m_P - m_F$, where m_P and m_F denote the criterion-means for those who pass and those who fail respectively). With a graded criterion this provides the quickest method of all. It can only be safely adopted when we have grounds for assuming that the standard deviation of the criterion-assessments (σ) remains constant for all items (as when all items are given to the same group of children), and that the differences in the number who pass or fail (p or q) will not seriously affect the results.

B. CRITERION DICHOTOMOUS (FOURFOLD FREQUENCY TABLES).

1. Correlational Measures.

(a) Item Assumed to follow a Normal Distribution.

(v) *The Normal Tetrachoric Correlation* (Pearson's r_t). According to Burt (11) this is the best procedure, when neither of the underlying distributions is likely to depart greatly from the normal. Even when the criterion-assessments are actually given in graded form, they may still be reduced to a twofold classification; and r_t may be used with advantage, not only because it saves time, but because, although it makes less use of the detailed information available, it nevertheless avoids the distorted values often given by r_{bis} when the distribution is not normal. Burt also points out that some of the information may be recovered, and the probable error reduced, by using the 'averaged tetrachoric'—i.e., averaging the coefficients obtained with different borderlines.

(b) Item Assumed to follow a Point-Distribution.

(vi) *The Product-Moment Correlation for a Fourfold Table* (Yule's ϕ). This would seem to be the appropriate coefficient when both twofold classifications are based on point-distributions.² Since $\phi^2 = \chi^2/N$, it possesses a readily available significance test.

(vii) *Yule's Coefficient of Colligation* (ω). This gives the point-correlation (ϕ) to be expected in an equalized fourfold table. When the actual numbers in the upper and lower criterion-groups are arbitrary, and we have no knowledge of their true numbers or true proportions, this seems the most appropriate coefficient. Yule argued, and Pearson strenuously denied, that it might be legitimately extended to more general cases. Note that, for $p = q = 0.50$, $r_t = \sin \frac{\pi}{2} \omega$. Whipple's Table 6 (*Manual of Mental and Physical Tests*, p. 36) may accordingly be used to effect the conversion.

¹ For the percentage measure of overlapping, and its reduction to terms of the standard deviation, see Burt, *J. Exp. Ped.*, I, 1912, p. 283; II, p. 253f., and *Distribution of Abilities*, 1917, p. 69. Vincent (*loc. cit. sup.*) preferred this device (viz., the percentage among the children failing in the item who exceed the median of the children passing the item): Barthelmess and Long have also adopted a modification of this general principle (5, pp. 430-1). A further refinement of this method can be used when the 'constant method' is employed for scaling items (see below).

² Pearson terms this coefficient the 'Boas-Yulean ϕ ': he himself prefers to use, not ϕ (the 'root mean square contingency'), but C_2 (the 'coefficient of mean square contingency,' where $C_2 = \sqrt{\phi^2/(1 + \phi^2)}$). Pearson has made several comparisons of the different methods of assessing association in the case of fourfold tables, and concluded: "the best method is first to investigate P (by calculating χ^2) and throw out not-associated (i.e., non-significant) cases; then to use either tetrachoric r_t or C_2 , according as we are justified in considering the variates continuous or not" (3, Vol. I, p. xxxvi; cf. Pearson, K., and Heron, D., 'On Theories of Association,' *Biometrika*, IX, 1913, pp. 167f.).

(2) Difference Measures.

(viii) *The Percentage Difference* ($d_p = p_U - p_L$, where p_U and p_L denote the percentages passing the item in the upper and lower criterion groups respectively). This is related to the point correlation coefficients in much the same way as the difference between the means is related to the normal correlation coefficients¹; and, like that, may be used as a simple time-saving substitute. It is the procedure that commonly first suggests itself to the non-statistical investigator. Its defect is that, when the test-items to be compared differ widely in difficulty, the omission of any correction for difference in total p and q is apt to give misleading results. When their difficulty remains much the same it generally gives the same order of efficiency as r_t .

Of the other methods available,² e.g., analysis of variance, factor analysis, etc., the one most useful for the Binet scale is the adaptation of the 'constant method' for scaling test-items: with this the 'threshold' expresses the degree of difficulty for the item, and so indicates its appropriate age-assignment, while the 'standard deviation' gives a measure of discriminative power, i.e., of validity (2, pp. 442-8). Terman and Merrill seem to have relied largely on this principle, since they took the 'steepness of the curves' (which might be measured numerically by the reciprocal of the standard deviation) as affording "a graphic indication of the validity of the tests" (6, p. 8).

Fine and Coarse Discrimination. Besides the nature of the distributions, the ultimate purpose of the test-scale must also be taken into account. Thus, if the object is to obtain a classification or set of measurements over a wide range of ability, then the 'heterogeneous' (or 'graded') type of test with items varying widely in difficulty will be most serviceable (2, p. 298). If, on the other hand, the object is to obtain a precise discrimination among a selected group (as in diagnosing borderline defectives or in the final examinations for junior county scholarships), then a 'homogeneous' (or 'uniform') type of test is more appropriate; and that implies that we are concerned with a test-scale (or part of a test-scale) in which all the items are of approximately the same level of difficulty (usually about 50 per cent. of the examinees should pass on each). For the former purpose r_{bis} or r_t would, in theory, provide the best validation formulæ, since they are (as we shall see) but little affected by differences in difficulty. For the latter purpose, however, ϕ , ω , η , and the percentage-difference may claim advantages since they favour items of medium difficulty. The difference-methods tend to give magnified values for items that are exceptionally difficult or exceptionally easy.

Criterion. In whatever form the assessments are expressed, the source of the criterion may be either 'external' (e.g., teachers' independent assessments) or 'internal' (i.e., based on the test as a whole—e.g., the total mark or 'mental age')

¹ It may be noted that $\frac{d}{\sigma} = \frac{m_p - m_F}{\sigma}$ expresses the slope of the regression of the criterion on the test-item, while $p_U - p_L$ expresses the slope of the regression of the test-item on the criterion (Burt, C., 'Statistics in Vocational Psychology,' *Occ. Psych.*, XVI, p. 172).

² In the foregoing account the formulæ used to define the several procedures are expressed in terms of the means (or percentages) of the two subgroups. For purposes of computation it is quicker to substitute for one of these values the mean (or percentage) for the entire group, since that figure remains constant for all items. With fourfold tables a simple device is to take the total number of disagreements (D): then $r_t = \cos 1.8 D^\circ$ (which can be read off from a quadrant). For these and other simplified working-methods, see 11. It should be remembered that, as Burt has pointed out, a factor analysis can be made for ϕ -coefficients, or even for the percentages on which they are based, just as well as for product-moment coefficients.

derived from all the items, or the 'factor-measurement' obtained by factorizing the intercorrelations for the items).

In studying the items in the Binet test, the criteria generally available are the following : (i) the teachers' grades or ranks for the children tested ; (ii) the teachers' classification as 'above' and 'below' some specified borderline (usually the average) ; (iii) a division of the sample into normal and mentally defective children ; (iv) a division into 'extremely bright' (e.g., scholarship winners or secondary school pupils) and 'extremely dull' (e.g., mentally defective or special school pupils), with the ordinary average children omitted ; (v) the graded assessments furnished by the mental ages or mental ratio (I.Q.) ; (vi) the test-assessments themselves, often reduced to a dichotomous or trichotomous classification. Now, as Burt (11) points out, when the test-scale itself forms the sole criterion, the resulting correlations provide a measure of 'consistency' or 'reliability,' rather than (as is so often assumed) of 'validity' in the technical sense.¹ On the other hand, teachers' assessments are apt to be biased in favour of scholastic aptitude rather than general intelligence as such, i.e., they are liable to attract excessive importance to industry, learning-capacity, and verbal and numerical as contrasted with practical performances. Hence the best available criterion would seem to be a combination of evidence obtained from both external and internal sources, the one confirming or correcting the other.

The Relative Suitability of the Different Procedures. All the methods described above can be used with test-items like those of the Binet scale ; and all have been given a trial in published or unpublished researches. In validating the items in the original version, Burt relied chiefly on the coefficient of colligation (ω), and published a graph or 'abac' for deducing the coefficient from the percentages of passes and failures (2, pp. 217, 231). He argued that it furnished much the same order of merit as methods assuming a normal distribution (tetrachoric or biserial r) ; was sufficiently reliable, if not more reliable, with small groups ; and was possibly the most appropriate in validating the items of the original Binet scale, since that was intended chiefly for discriminating borderline defectives. But, in general, he considered, "where an 'internal grading' (for the item-responses) and a full product-moment formula are out of the question, then for final conclusions tetrachoric r should be calculated at length, and other formulæ used as controls" (2, p. 220).

In validating the items of the Stanford Revision as applied to English children the tetrachoric correlation was preferred : once again graphs were prepared for the purpose, based on Pearson's tables (3). For validating items from written group-tests, where large numbers can be tested, and fairly smooth normal distributions can be secured, biserial r appeared to yield the most satisfactory figure. For individual tests, however, it proved to be less trustworthy, for the reasons given above.

In all these cases the selection of this or that procedure seems to have been determined chiefly by general impression or by *a priori* argument. More recently, however, a number of investigations have been carried out, particularly in America, with the object of measuring the validity of the validity-coefficients themselves. The general procedure has been to correlate the indices obtained by the different

¹ The argument is that the correlation of an item with the summed marks for all the items is proportional to the sum of the correlations with the marks for the separate items. It therefore represents the sum of the intercorrelations for that item. If, however, the real object of the investigation is deliberately to study 'reliability' or 'consistency,' an equivalent but more effective method is to use analysis of variance (cf. Burt, C., *Brit. J. Educ. Psych.*, XV, pp. 87f., section on 'Reliability of Pass-or-Fail Tests'). However, in most attempts at test-construction American investigators still seem to rely by preference on the total score for the criterion. With the Terman-Merrill scale, the 'item-correlations with total score' as obtained from American children have recently been published (8, pp. 175f.) ; and it is stated that "the retention of an item depended in part upon its correlation with the total score."

methods (a) with each other, (b) with some standard method as a criterion (e.g., the full product-moment method), or (c) with disturbing factors known to affect the calculated value of the coefficient (e.g., the ease or difficulty of the item or total percentage passing it).¹ No such investigation, however, seems to have been published, specifically for the items of the Binet scale. We have therefore to rely mainly on a preliminary enquiry made by Miss Matthews and others for the Binet Committee.

As her standard she took the figures obtained by the full product-moment method: in the later versions of the scale a number of items can be made to furnish graded assessments; and, if the criterion-assessments are also in graded form, the correlation can then be calculated in the ordinary way (cf. 2, p. 220; 8, p. 122). Following Miss Swineford,² correlations were also calculated between the various indices and with the ease or difficulty of the test for the group in question as indicated by the number of those passing or failing (i.e., with $|\frac{1}{2} - p|$, which may be taken as measuring the degree to which the line of division is shifted from the median). The figures are reproduced in Table I. As Miss Matthews and her colleagues point out, "the table of correlations suggests a general factor, which is evidently the validity of the item, and a bipolar factor, which seems to distinguish methods based on normal or graded assessments from those based on percentages and point-distributions. In fact the three percentage-methods seem evidently inappropriate in the case of a quality which, like intelligence, is normally distributed. Similarly both the simple difference-methods seem relatively unreliable.³ We are left with r_{bis} and r_t . On the whole, provided the lines of division do not fall too far from the median, and a watch is kept for erratic distribution, the tetrachoric correlation would appear to give the best and the quickest all-round results. The biserial method may be safer with very small groups, though these are apt to yield erratic results; and the point-distributions will be suitable with extreme, exceptional, or arbitrary selections of the two contrasted groups."⁴

TABLE I. CORRELATIONS BETWEEN VALIDITY-COEFFICIENTS FOR BINET TESTS

Method	d_m	r_{bis}	r_t	ω	φ	p_m
Mean-Difference (d_m)	—	.97	.92	.71	.68	.53
Biserial Correlation (r_{bis})	.97	—	.96	.67	.76	.65
Tetrachoric Correlation (r_t)	.92	.96	—	.83	.82	.69
Coefficient of Colligation (ω)	.71	.67	.83	—	.91	.84
Point Correlation (φ)	.68	.76	.82	.91	—	.95
Percentage-Difference (p_m)	.53	.65	.69	.84	.95	—
General Factor Saturations	.88	.92	.96	.89	.94	.84
Bipolar Factor Saturations	-.42	-.37	-.22	.24	.33	.44
Product-Moment Correlation (r)	.69	.93	.98	.85	.76	.61
Correlation with Dichotomy ($ \frac{1}{2} \pm p $)	-.25	-.12	-.15	.09	.31	.43

¹ For reviews of such studies see 5, and more particularly 4 which is not included in Guilford's bibliography (5, pp. 453-6).

² Swineford, F., 'The Validity of Test Items,' *J. Educ. Psych.*, XXVII, 1936, Table I, p. 73.

³ In the memorandum d_m is used to denote the *standardized* difference, i.e., $(m_P - m_F)/\sigma$.

⁴ It is emphasized that these conclusions are only meant to apply to intelligence tests of the Binet type. For other qualities and for other types of test, other methods may be preferable: thus, for group tests, where a rough and rapid assessment for numerous items obtained from large populations is needed, the crude difference methods may be sufficient.

Since her own groups were rather small, it was decided to verify her main conclusions with the larger groups now available. Graded assessments were available for Vocabulary (three age-groups), Repeating Numbers (two age-groups), Picture Vocabulary, Naming Words, Abstract Words, and Opposites (one age-group each). Nine product-moment correlations could therefore be calculated from these tests; and the correlations themselves were correlated with (a) the tetrachoric correlations, (b) the biserial correlations, and (c) the colligation coefficients for the same tests and age-groups. The figure so obtained expresses the agreement of the several shortened methods with the full product-moment method. The correlation was 0.89 for the tetrachoric procedure, 0.81 for the biserial, and 0.73 for the colligation coefficients. By averaging three determinations of the tetrachoric coefficient the figure was raised to 0.94.

Present Procedure. Accordingly, for the present data the tetrachoric correlation was chosen as the standard procedure. The criterion was the teachers' assessments, checked by the mental age of the children tested; or, when no assessments were available from the teachers, the mental age itself. On this basis the groups were divided into two subgroups approximately equal in numbers, namely, those above and those below the median.¹ As a further check two supplementary calculations were made for the largest groups of all with the divisions at the upper and lower quartiles respectively; and for these a biserial coefficient was also computed. In most cases these and other alternative determinations that were tried furnished much the same order of validity, though the absolute values sometimes varied as much as ± 0.15 . The figures finally adopted have been based on the first of these methods alone.

III. RESULTS

General Results. The figures thus obtained are shown in Table II. The last column but one gives the revised age-assignments for each test, calculated by the procedure described for the earlier revision (2, pp. 150-1, 442-8). And the last column gives the validity coefficients. These range from .21 (Block-building, age 2) to .78 (Vocabulary, age 14). None of the correlations approaches the highest values reported with American children. For the whole series the average coefficient is .51. For ages 3 to 16 it is .53, as compared with .45 for the original Binet and .49 for the Stanford Revision.

Results for the Several Tests. For those items which figured in earlier versions of the scale the coefficients here obtained agree tolerably well with those previously published for English children. On the whole, too, our figures correspond as closely as could be expected with those obtained from American children (9). There are

¹ With subjective assessments it is very much a matter of chance to which side of the median line children of roughly average ability get assigned. For this reason Burt has argued that the best line of division will not necessarily be at the median. With ω , he suggests that, when the test divides the group into two unequal portions, p and q , the criterion should be divided in the same way (2, p. 206). With r_n , where the coefficient itself is assumed to correct for the unequal division, all divisions should give the same results; and hence (particularly with small samples) it will be well to try different lines of division, and, if the results vary, to take a weighted average. The best single point of division he claims (10, p. 222) will be at $x = \frac{1}{2} \bar{x}_t$ (where $\bar{x}_t$ is the mean deviation of the tail so cut off): this gives $p = 27.03$ per cent. or approximately 25 per cent. However, where objective assessments are available in place of, or as a check on, the subjective ratings, the data here analysed would seem to indicate that the best borderline is after all at the median. But the problem calls for further investigation with the aid of actual data rather than of a *a priori* argument.

American investigators, however, appear strongly to favour biserial r (5, p. 431). Merrill and McNemar have preferred this procedure. But some of their figures appear extremely high for isolated items: 38 of the 134 items have coefficients of .70 or over; seven over .80 (9, pp. 176-180). Miss Matthews, in her preliminary item-analysis, had previously reported that "the biserial coefficient seems at times to yield grossly exaggerated figures and grossly exaggerated differences."

I am much indebted for assistance with the preliminary calculations to Miss Bruce and Miss Banks. A fuller comparison of the results will be found in an unpublished thesis dealing with the subject.

a few notable exceptions, e.g., for Memory Drawing the American data yield a figure of only .28—one of the lowest in the list.

So far as the results of factorial analyses are available for comparison, the order of validity given in Table II appears to agree with the order given by previous writers for the saturations for the general factor.¹ Thus, what earlier observers found for the earlier scale holds good of this : “ the order of efficiency for the tests as given by the saturation coefficients (obtained by factor analysis with an *internal* criterion) is much the same as that obtained by the method of item-analysis with an independent (or *external*) criterion : . . . judged by these figures the best would seem to be—absurdities, mixed sentences, vocabulary, abstract words, memory-drawing, and repeating numbers backwards ” (8, p. 127). Here the order of efficiency for the eight best tests is—vocabulary, abstract words, sentence-building, similarities and differences, analogies, sentence-completion, verbal absurdities, and reasoning (memory-drawing and repeating numbers backwards do not come quite so high in the list as before).

The tests devised for children of elementary school age prove on the whole to have the highest validity (average .54) ; those for older children and young people beyond the school leaving age are not quite so effective (average .51) ; and those for pre-school children least effective of all (average .38). The same type of test, however, may have a very different value at different age-levels : the validity of the vocabulary test, for example, increases steadily throughout the school period and then tends to decline.

On the whole, the verbal tests (average .53) have a higher validity than the pictorial or manipulative (averages .35 and .43 respectively). Some of the attractive non-verbal tests introduced by Terman and Merrill (e.g., the Bead and Block-building tests) prove decidedly disappointing. Their low validity can scarcely be ascribed to the excess of verbal tests in the criterion or to the teachers' scholastic bias, since it holds good even in groups for which an assessment based on out-of-school efficiency was available (Boy Scouts and Girl Guides).

In view of Spearman's criticisms of the scale, it is instructive to re-arrange the several test-items according to McDougall's hierarchy of cognitive levels. Those tests in the original version which involved little else but simple sensory discrimination, and gave the lowest validity-coefficients of all, have already been dropped ; among the tests that remain, those depending mainly on simple perception or the most elementary forms of motor skill yield the lowest values ; tests depending on mechanical memory give slightly higher figures ; the highest coefficients of all are given by tests involving the highest type of mental processes, the perception and application of relations, particularly problems involving generalization, abstraction, the use of analogy, and critical or constructive reasoning. This agrees with the results obtained with other types of intelligence test (e.g., graded tests and group tests) : “ The higher the mental level tested and the more complex the process involved, the more completely do the results correspond with independent estimates of intelligence. Tests of the highest processes of all, those involving relations, and usually classed together as ‘ reasoning,’ yield the highest correlations.”²

¹ Cf. Miss Davidson's preliminary study of the Terman-Merrill results, briefly reported in 8, p. 118.

² Burt, C., ‘ Tests of Higher Mental Processes and their Relation to General Intelligence ’ (*J. Exp. Ped.*, I, 1911, pp. 95, 103, 112, and other papers). This conclusion was put forward in opposition to Spearman's claim that ‘ general intelligence ’ and ‘ general sensory discrimination ’ are identical (a claim perhaps suggested more by the earlier view of British writers like Bain and Sully, who regarded discrimination as an essential process in intelligence, than by actual experimental proof). It is a little surprising therefore to find American writers attributing the partial or complete identification of the ‘ general factor ’ with ‘ reasoning ability ’ to Holzinger or Spearman (cf. Wolfe, *Factor Analysis* to 1940, p. 33). Spearman did not accept such a view (or something approaching it) till 1921. Spearman's later assertion that Binet “ borrowed the idea ” of heterogeneous testing from Spearman's 1904 article (*Abilities of Man*, 1926, pp. 71 and 72) is quite unjustified : see Binet's own criticisms of that article at the time (*Ann. Psych.*, XI, 1905, pp. 624–5).

The fact that a test-item has a low correlation with the general factor of intelligence does not necessarily imply that it should be discarded forthwith. First of all, it may throw suggestive light on special abilities or disabilities. Binet's original intention, it may be remembered, was that his tests might indicate the development of particular capacities or 'faculties' as well as general intelligence; and in actual practice it is often convenient to include tests which may reveal weakness in some special 'group factor' such as mechanical memory or numerical ability. This however, is a question to be discussed when fuller information is available from the results of factor analysis.¹ Secondly, the best composite test-scale will be obtained, not by merely selecting those items which correlate most highly with the criterion, but by constructing a well-balanced battery of items having low correlations with each other: thus non-verbal items may require to be retained, despite their low validity, because they usefully counterbalance the verbal tendencies of the more reliable tests (see 11, pp. 19f.). What has been called the technique of 'item-synthesis, is therefore required to supplement that of 'item-analysis.'² Finally, it must be remembered that a thorough study of the value of each item involves much more than the simple determination of its 'validity' in the technical sense. As we saw at the outset, its ultimate suitability can only be judged after a number of supplementary determinations, qualitative as well as quantitative,³ have been carried out. Hence the foregoing enquiry must be taken as dealing only with one important aspect of the total problem.

IV. SUMMARY AND CONCLUSIONS

1. Previous item-analyses for earlier versions of the Binet scale have shown that test-problems presumed to have a high validity in the country where they were first proposed often show a much lower validity when given to children in this country. Hence it is argued that the Terman-Merrill revision cannot be safely adopted until the value of the separate items has been checked by application to British children.

2. Test-results are now available for over 2,000 children and young persons, roughly 150-200 per age-group among those of elementary school age. The criterion adopted has been the teachers' assessments of intelligence, checked by the results of the tests as a whole, or, where no teachers' estimates were available, the test-results themselves.

3. The procedures available for estimating item-validity have been briefly reviewed. It is argued on theoretical grounds that the tetrachoric correlation, with the lines of division as near the median as possible, should be the most satisfactory for the present purpose. Validity-coefficients obtained in this way have been compared with those obtained by alternative methods, and the empirical results conform with this expectation.

4. The correlation coefficients and revised age-assignments for each test-item have been given in Table II. Taken as a whole the items in the new revision appear slightly more efficient than those of earlier versions. But the validity of the new non-verbal tests and particularly of the tests for pre-school years is decidedly inferior to that of some of the earlier types of verbal test, particularly those devised for children of school age. The most effective are the tests of the higher mental processes (such as those involving the perception and application of relations), e.g., abstraction, generalization, and reasoning.

¹ For preliminary results see 7. The method of 'correlating persons' (both examinees and examiners) which was used with fruitful results for the earlier scale might also be applied to the data now available for the newer scale (see 7, pp. 257-260).

² *Loc. cit. sup.*, *Brit. J. Educ. Psych.*, XV, p. 87.

³ See, for example, Burt's method of studying items for the Opposites Test, which he offers "as an illustration of the full procedure and a guide to those who wish to attempt the construction and standardization of other tests for themselves" (2, p. 235).

TABLE II. ORDER OF DIFFICULTY AND VALIDITY-COEFFICIENTS FOR BINET TESTS (TERMAN-MERRILL REVISION)

AGE 2								Difficulty (in Mental Age)	Validity- Coefficient
1.	Identifying parts of the body (3 points) ..	..	..	..	..	..	..	1.7	.39
2.	Identifying objects by name (3 points) ..	..	..	..	..	..	..	1.8	.42
3.	Three-hole form-board (1 point) ..	..	..	..	..	..	..	1.8	.36
4.	Obedying instructions (2 points) ..	..	..	..	..	..	..	1.9	.27
5.	Block-building : tower ..	..	..	..	..	..	..	2.0	.21
6.	Word combinations. (Alternative only) ..	..	..	..	..	..	..	2.0	.34
Average ..									.33
AGE 2½									
1.	Picture vocabulary (2 points) ..	..	..	..	..	..	..	2.2	.41
2.	Identifying parts of the body (4 points) ..	..	..	..	..	..	..	2.2	.34
3.	Rotated form-board (1 point) ..	..	..	..	..	..	..	2.4	.28
4.	Identifying objects by use (3 points) ..	..	..	..	..	..	..	2.4	.36
5.	Identifying objects by name (5 points) ..	..	..	..	..	..	..	2.5	.39
Average ..									.36
AGE 3									
1.	Naming objects ..	..	..	..	..	..	..	2.6	.30
2.	Repeating two numbers ..	..	..	..	..	..	..	2.7	.38
3.	Block-building : bridge ..	..	..	..	..	..	..	2.8	.35
4.	Picture vocabulary (9 points) ..	..	..	..	..	..	..	2.9	.51
5.	Stringing beads ..	..	..	..	..	..	..	3.0	.18
6.	Rotated form-board (2 points). (Alternative only) ..	..	..	..	..	..	..	3.0	.33
Average ..									.34
AGE 3½									
1.	Comparing sticks ..	..	..	..	..	..	..	3.1	.42
2.	Picture memories. (Alternative only) ..	..	..	..	..	..	..	3.3	.39
3.	Obedying instructions (3 points) ..	..	..	..	..	..	..	3.3	.46
4.	Picture vocabulary (12 points) ..	..	..	..	..	..	..	3.4	.41
5.	Copying a circle ..	..	..	..	..	..	..	3.5	.27
Average ..									.39
AGE 4									
1.	Repeating three numbers ..	..	..	..	..	..	..	3.5	.43
2.	Copying a cross. (Alternative only) ..	..	..	..	..	..	..	3.5	.26
3.	Describing pictures ..	..	..	..	..	..	..	3.6	.37
4.	Picture vocabulary (15 points) ..	..	..	..	..	..	..	3.7	.49
5.	Discriminating forms ..	..	..	..	..	..	..	3.7	.53
6.	Counting four objects ..	..	..	..	..	..	..	3.8	.45
7.	Comparing faces ..	..	..	..	..	..	..	3.9	.35
8.	Identifying objects by use (5 points) ..	..	..	..	..	..	..	4.0	.44
9.	Comprehension, I ..	..	..	..	..	..	..	4.0	.43
Average ..									.42

An Item-Analysis of the Terman-Merrill Revision

AGE 4½								Difficulty (in Mental Age)	Validity- Coefficient
1.	Picture vocabulary (16 points)	..	..	..	..	..	..	4·1	·61
2.	Naming objects from memory	..	..	..	..	..	..	4·2	·52
3.	Pictorial identification (3 points)	..	..	..	..	..	..	4·2	·38
4.	Picture completion (1 point)	..	..	..	..	..	..	4·3	·43
5.	Comprehension, II	..	..	..	..	..	..	4·4	·54
6.	Repeating 11 syllables. (Alternative only)	..	..	..	..	..	..	4·5	·45
Average ..									·49

AGE 5								Difficulty (in Mental Age)	Validity- Coefficient
1.	Copying a square	..	..	..	..	..	..	4·5	·33
2.	Triple order	..	..	..	..	..	..	4·5	·45
3.	Picture completion (2 points)	..	..	..	..	..	..	4·6	·46
4.	Pictorial identification (4 points)	..	..	..	..	..	..	4·6	·42
5.	Definitions	..	..	..	..	..	..	4·7	·38
6.	Paper folding (triangle)	..	..	..	..	..	..	4·9	·35
7.	Repeating 12 syllables	..	..	..	..	..	..	4·9	·47
8.	Repeating four numbers	..	..	..	..	..	..	5·0	·56
Average ..									·43

AGE 6								Difficulty (in Mental Age)	Validity- Coefficient
1.	Vocabulary test (five words)	..	..	..	..	..	..	5·7	·58
2.	Opposite analogies (2 points)	..	..	..	..	..	..	5·2	·63
3.	Materials	..	..	..	..	..	..	5·2	·55
4.	Number concepts	..	..	..	..	..	..	5·3	·49
5.	Copying a diamond	..	..	..	..	..	..	5·3	·41
6.	Making a knot	..	..	..	..	..	..	5·5	·34
7.	Copying a bead chain from memory	..	..	..	..	..	..	5·6	·27
8.	Repeating five numbers	..	..	..	..	..	..	5·8	·51
9.	Pictorial likenesses and differences (5 points)	..	..	..	..	..	..	5·9	·57
Average ..									·48

AGE 7								Difficulty (in Mental Age)	Validity- Coefficient
1.	Vocabulary (seven words)	..	..	..	..	..	..	6·6	·61
2.	Repeating five numbers	..	..	..	..	..	..	6·0	·47
3.	Tracing a maze	..	..	..	..	..	..	6·1	·51
4.	Repeating 16 syllables	..	..	..	..	..	..	6·2	·55
5.	Picture absurdities, I	..	..	..	..	..	..	6·2	·44
6.	Mutilated pictures	..	..	..	..	..	..	6·3	·51
7.	Comprehension, III	..	..	..	..	..	..	6·7	·49
8.	Repeating three numbers backwards	..	..	..	..	..	..	6·9	·57
Average ..									·52

AGE 8								Difficulty (in Mental Age)	Validity- Coefficient
1.	Vocabulary (nine words)	..	..	..	..	..	..	7·4	·63
2.	Opposite analogies (5 points)	..	..	..	..	..	..	7·2	·58
3.	Similarities between two things	..	..	..	..	..	..	7·3	·61
4.	Memory for stories ("A Fall in the Mud")	..	..	..	..	..	..	7·5	·55
5.	Comprehension, IV	..	..	..	..	..	..	7·7	·52
6.	Verbal absurdities, I (3 points)	..	..	..	..	..	..	8·0	·57
Average ..									·58

AGE 9								Difficulty (in Mental Age)	Validity- Coefficient
1.	Vocabulary (11 words)	..	..	..	..	..	..	8.5	.66
2.	Giving change	..	..	..	..	..	..	8.1	.53
3.	Repeating six numbers	..	..	..	..	..	..	8.1	.42
4.	Similarities and differences	..	..	..	..	..	..	8.3	.64
5.	Repeating four numbers backwards	..	..	..	..	..	..	8.4	.47
6.	Memory for two designs (1 point)	..	..	..	..	..	..	8.9	.45
7.	Verbal absurdities, II (2 points)	..	..	..	..	..	..	9.0	.58
Average		..	..	..	..	..	..		.54

AGE 10									
1.	Vocabulary (13 words)	..	..	..	..	..	..	9.4	.65
2.	Paper cutting (1 point)	..	..	..	..	..	..	9.1	.63
3.	Naming words (28 points)	..	..	..	..	..	..	9.1	.47
4.	Restricted rhymes	..	..	..	..	..	..	9.2	.51
5.	Picture absurdities ("Frontier Days")	..	..	..	..	..	..	9.5	.55
6.	Memory for two designs (1½ points)	..	..	..	..	..	..	9.5	.53
7.	Reading and report (10 points)	..	..	..	..	..	..	9.7	.48
8.	Repeating 20 syllables	..	..	..	..	..	..	9.8	.45
Average		..	..	..	..	..	..		.53

AGE 11									
1.	Vocabulary (15 words)	..	..	..	..	..	..	10.6	.71
2.	Verbal absurdities, III (2 points)	..	..	..	..	..	..	10.1	.62
3.	Finding reasons, I	..	..	..	..	..	..	10.4	.52
4.	Abstract words, set I (3 points)	..	..	..	..	..	..	10.4	.65
5.	Sentence building with three words	..	..	..	..	..	..	10.7	.68
6.	Repeating five numbers backwards	..	..	..	..	..	..	10.7	.49
7.	Answering difficult questions. (Alternative only)	..	..	..	..	..	..	11.0	.46
Average ..		..	..	..	..	..	..		.59

AGE 12									
1.	Vocabulary (17 words)	..	..	..	..	..	..	11.5	.68
2.	Mixed sentences	..	..	..	..	..	..	11.2	.59
3.	Abstract words, set II (2 points)	..	..	..	..	..	..	11.3	.63
4.	Purse and field	..	..	..	..	..	..	11.3	.48
5.	Memory for words	..	..	..	..	..	..	11.6	.56
6.	Response to pictures, II	..	..	..	..	..	..	11.7	.52
7.	Similarities between three things	..	..	..	..	..	..	11.7	.57
8.	Picture absurdity ("The Shadow")	..	..	..	..	..	..	11.9	.55
Average ..		..	..	..	..	..	..		.57

AGE 13									
1.	Vocabulary (19 words)	..	..	..	..	..	..	12.5	.74
2.	Completing sentences (2 points)	..	..	..	..	..	..	12.1	.62
3.	Abstract words, set II (3 points)	..	..	..	..	..	..	12.2	.73
4.	Induction	..	..	..	..	..	..	12.5	.60
5.	Copying a bead chain from memory, II	..	..	..	..	..	..	12.6	.43
6.	Repeating seven numbers	..	..	..	..	..	..	12.8	.48
Average ..		..	..	..	..	..	..		.60

An Item-Analysis of the Terman-Merrill Revision

AGE 14								Difficulty (in Mental Age)	Validity- Coefficient
1. Vocabulary (21 words)	..	..	..	..	..	..	..	13.3	.78
2. Paper cutting, I (2 points)	..	..	..	..	..	..	..	13.2	.53
3. Orientation, I	..	..	..	..	..	..	..	13.5	.44
4. Problems of fact	..	..	..	..	..	..	..	13.7	.32
5. Ingenuity, I (1 point)	..	..	..	..	..	..	..	13.7	.57
Average	..	..	..	..	..	..	..		.53
AVERAGE ADULT (AGE 15 $\frac{4}{12}$)									
1. Vocabulary (23 words)	..	..	..	..	..	..	..	15.1	.68
2. Arithmetical reasoning	..	..	..	..	..	..	..	14.2	.57
3. Repeating 25 syllables	..	..	..	..	..	..	..	14.6	.49
4. Enclosed boxes	..	..	..	..	..	..	..	14.7	.33
5. Codes	..	..	..	..	..	..	..	15.0	.38
6. Essential similarities	..	..	..	..	..	..	..	15.2	.51
7. Repeating six numbers backwards	..	..	..	..	..	..	..	15.3	.46
Average	..	..	..	..	..	..	..		.49
SUPERIOR ADULT, I (AGE 17 $\frac{4}{12}$)									
1. Vocabulary (27 words)	..	..	..	..	..	..	..	16.7	.54
2. Differences between abstract words	..	..	..	..	..	..	..	15.8	.63
3. Proverbs, I	..	..	..	..	..	..	..	16.1	.59
4. Completing sentences (3 points)	..	..	..	..	..	..	..	16.3	.53
5. Reconciling opposites (3 points)	..	..	..	..	..	..	..	17.1	.46
6. Ingenuity, II (3 points)	..	..	..	..	..	..	..	17.3	.57
Average	..	..	..	..	..	..	..		.55
SUPERIOR ADULT, II (AGE 19 $\frac{10}{12}$)									
1. Vocabulary (32 words)	..	..	..	..	..	..	..	18.6	.46
2. Reconciling opposites (5 points)	..	..	..	..	..	..	..	17.5	.47
3. Summarizing passage: "The Value of Life"	..	..	..	..	..	..	..	17.8	.52
4. Proverbs, II	..	..	..	..	..	..	..	18.5	.63
5. Orientation, II	..	..	..	..	..	..	..	19.0	.38
6. Paper cutting, II	..	..	..	..	..	..	..	19.5	.41
7. Repeating eight numbers	..	..	..	..	..	..	..	19.8	.49
Average	..	..	..	..	..	..	..		.50
SUPERIOR ADULT, III (AGE 22 $\frac{10}{12}$)									
1. Vocabulary (38 words)	..	..	..	..	..	..	..	21.5	.52
2. Finding reasons, II	..	..	..	..	..	..	..	20.0	.36
3. Opposite analogies	..	..	..	..	..	..	..	20.5	.63
4. Reasoning	..	..	..	..	..	..	..	22.0	.59
5. Repeating nine numbers	..	..	..	..	..	..	..	22.8	.48
Average	..	..	..	..	..	..	..		.52
General Average	..	..	..	..	..	..	..		.49

Note.—The 'mental ages' for tests assigned to 'adult' levels are conventional values only and are intended to indicate the approximate increments of difficulty on the same scale as the earlier age.

REFERENCES

1. Burt, C. 'The measurement of intelligence by the Binet tests,' *Eugenics Review*, VI, 1914, 36-50, 140-152.
2. Burt, C. *Mental and Scholastic Tests* (2nd ed., 1947). London : Staples Press, 1921.
3. Pearson, K. *Tables for Biometricians*, Vol. II. London : Biometric Laboratory, University College, 1931.
4. Long, J. A., and Sandiford, P. *The Validation of Test Items*. Toronto : Dept. Educ. Res. (Bulletin 3), 1935.
5. Guilford, J. P. *Psychometric Methods*. New York : McGraw-Hill Book Company, 1936.
6. Terman, L. M., and Merrill, M. A. *Measuring Intelligence*. London : Geo. Harrap and Co., 1937.
7. Burt, C. The latest revision of the Binet intelligence tests. *Eugenics Review*, XXX, 1939, 255-260.
8. Burt, C., and John, E. 'A factorial analysis of the Terman-Binet tests,' *Brit. J. Educ. Psych.*, XII, 1942, 156-161.
9. McNemar, Q. *The Revision of the Stanford Binet Scale*. New York : Houghton Mifflin Co., 1942.
10. Burt, C. 'Statistical problems in the evaluation of Army tests,' *Psychometrika*, IX, 1944, 219-235.
11. Burt, C. Item analysis in educational tests. 1945. (Unpublished memo.)
12. Burt, C. *Handbook of Tests for Use in Schools* (2nd ed.). London : Staples Press, 1948.

INDICES OF ITEM CONSISTENCY AND VALIDITY¹

By P. E. VERNON

Senior Psychologist's Department, The Admiralty

I. Introduction. II. Definitions. III. Grouping Methods. IV. Distribution Methods. V. Other Methods. VI. Discussion : (a) Speed and Reliability ; (b) Multiple Criteria ; (c) Detailed Analysis ; (d) Guessing Correction ; (e) Spurious Internal Consistency Correlations ; (f) Effects of Insufficient Time ; (g) Effects of Balance ; (h) Selection of Items for Maximum Validity. VII. Experimental Studies : (a) Tabulation Time ; (b) Inter-correlations and Correlations with Balance ; (c) Reliability. VIII. Summary and Conclusions.

I. INTRODUCTION

Hawkes and Lindquist (1937), Flanagan (1939), Davis (1946*a*) and others insist that the main factors in the construction of items for paper-and-pencil tests of aptitudes, attainments or personality qualities are the ingenuity and experience of the authors and editors. Each item must be carefully chosen and well-written, and scrutinized by experts to ensure that the correct answer is really correct, and incorrect alternatives really incorrect. No amount of statistical analysis can make up for weaknesses at this stage. At the same time professional test constructors never expect all their items to be satisfactory. Hence they normally devise an excess of items and employ item-analysis both to reveal unsuspected weaknesses which may be overcome by revision, and to eliminate the poorest items.

Though there has been much investigation of indices of satisfactoriness, most of the published literature is American.² The fundamental principles were indeed established by Burt (1921) in his analysis and revision of the Binet-Simon scale, and his construction of group intelligence tests. But since then, more attention has been paid by British psychologists to item-difficulty and its effects on test score distribution and reliability (Walker, 1931-40 ; Lawley, 1943) than to item-discrimination. However with the adoption by the Defence Services and the Civil Service Commission of large-scale psychological testing, the problem of the most suitable indices has come to the fore, and it has been possible to carry out relevant experimental investigations. The object of this article, then, is to review the techniques of analysis, and to report researches undertaken by the writer and his colleagues in Government employment.

II. DEFINITIONS

The following symbols will be employed throughout. The first seven were suggested by Anstey in the Army Directorate for Selection of Personnel, and are regularly used by the Civil Service Commission's Research Unit.

¹ This article is published with the approval of the Lords Commissioners of the Admiralty, but the responsibility for any statements of fact or opinions expressed rests solely with the author.

² Since this was written the author has been privileged to see two unpublished memoranda which provide valuable surveys of the whole field of item analysis, namely : " Item Analysis in Educational Tests," by Cyril Burt, and " The Construction and Item Analysis of Aptitude Tests " by E. Anstey. The conclusions of these writers are generally very similar to those put forward in this article.

R = the number of testees giving the right answer to an item. The terms right and wrong refer to tests of aptitudes and attainments. In the case of personality or interest questionnaires, attitude scales and the like, precisely the same techniques are applicable, but a 'right' response is one presumed to measure the desired quality, interest or attitude.

E = the number of testees giving an erroneous response to an item.

O = the number of testees reaching an item but omitting it, as shown by their attempting subsequent items.

$F = E + O$ = the number failing to do an item.

U = the number not reaching or attempting an item. When the time limit is sufficient for all to finish, U is zero and $W = F$.

$W = F + U$ = the total number getting an item wrong or omitting it.

$T = R + F$ = total number attempting an item.

$N = R + W$ = total size of the population whose responses are being analysed.

Summarising :

R = Right answer

E = Tried but wrong

O = Reached but omitted

U = Not reached

$$\left. \begin{array}{l} F = \text{Failed} \\ W = \text{Wrong} \end{array} \right\} T = \text{Tried} \left. \begin{array}{l} \\ \end{array} \right\} N$$

It should be realised that accurate classification of testees' responses under these categories is not always possible. Thus several apparent U 's should often be O 's, since testees may have attempted an item but recorded no response. On the other hand some O 's may actually be U 's, for testees often omit an early item without attempting it properly and go on to later items.

M_r, M_w , etc. = the mean criterion scores of the R, W , etc. groups.

$d = M_r - M_w$.

$p = R/N$ = the proportion of testees giving the right answer. $q = 1 - p$.

p_U, p_L = the proportions of subgroups giving right answers, U and L being upper and lower groups on the criterion.

σ = the standard deviation of criterion scores of the whole group.

σ_I = the standard deviation of scores on an item = $\sqrt{pq}$.

n = the number of alternatives in a selective-response item.

$z, C.R., t, F, \eta, r, r_{ts}, r_{biss}, \phi, \chi^2$, etc. bear their usual connotations.

Balance is Swineford's (1936a) term to indicate the closeness of item-difficulty to 50 per cent. in the group whose responses are being analysed. *Imbalance* is used to indicate that p approaches 1.0 or 0.

Criterion. The comparison of items with an *external* criterion, such as subsequent success of the testees at school examinations, or in a job, provides an index of their *validity*. Other examples of external criteria are chronological age (in the case of age-scales for intelligence, reading ability, etc.), or membership of distinctive groups such as mechanics vs. clerks (in a test of mechanical ability), and manic-depressives vs. schizophrenics (in a test of cycloid-schizoid traits). Comparison with an *internal* criterion, namely the score on the test itself, provides indices of item *consistency*.

Scores on other tests may be used as criteria intermediate between the external and internal types. Thus items for a new intelligence test can be compared with a well-established test or battery. Again a test may be devised to measure, or to avoid measuring, a given factor. Thus the items in the Army mechanical comprehension test were chosen to differentiate successful and unsuccessful mechanics, but were also analysed against g and v tests, in order that these factors could be held constant (Vernon and Parry, 1948).

Discrimination refers either to validity or consistency, i.e. to the correlation between an item and any chosen criterion.

III. GROUPING METHODS

Useful surveys of some twenty indices of validity or consistency are provided by Long and Sandiford (1935) and Guilford (1936a); (cf. also Lentz and Hirshstein, 1932; Lindquist and Cook, 1933; Zubin, 1934; Swineford, 1936a; Chapanis, 1941). Several of the rather cruder ones, which do not seem to have been applied in any recent researches, will be omitted here. Two main types of method may be distinguished, which we shall call Type *A* or Grouping methods, and Type *B* or Distribution methods. In the first the criterion scores are divided into two or more categories, while in the second they are treated as a continuous distribution. The first involves taking the answer sheet of each testee in a given category and tabulating his right (or other) responses to each item, before going on to the next testee. Thus it yields the numbers of right (or other) responses to each item by members of the several categories. The second involves taking each item in turn and tabulating the criterion scores of all testees who pass, and of those who give other responses to, that item before proceeding to the next one.

If the criterion is itself categorical, grouping methods are obviously necessary. But when it is continuous the groups may consist of :

- (i) testees above and below the median on the criterion ;
- (ii) the tails of the distribution, e.g. the highest and lowest quartiles or 10 per cents. The highest and lowest 27 per cents. are most frequently chosen in America since, as Kelley (1939) has shown these proportions yield the largest critical ratios between p_U and p_L (provided that criterion scores are normally distributed). It is generally assumed that the difficulty of the item can be adequately estimated from the mean of p_U and p_L , without tabulating the responses of the middle group. Obviously this assumption is unsafe if the tails are very small.

- (iii) Three groups (Highs, Mediums and Lows) are favoured by the present writer. Quartile and even decile groups have been used by others, and Thomson (1942) generally prefers six. In general each group should contain at least 50, preferably 100 testees, for reliable percentages to be obtained.

The groups are usually equal in size, or approximately so, to facilitate calculations, but need not be. Thus for an age scale the numbers of children available at different age levels may differ considerably. Provided the numbers of testees are large, it is an advantage to have several groups, since the discriminations of an item may vary considerably at different points on the scale (Hawkes and Lindquist, 1937). An easy item may be almost useless for differentiating above-average from below-average testees, and yet discriminate quite well between the bottom and the upper five sextiles. Zubin (1934) indeed suggests that the total group should be split for each item at the q^{th} percentile level (corresponding to the difficulty of that item). But this would enormously increase the complexity of tabulation.

We may now list the chief methods of treatment of the tabulated data.

1. *Graphical.* If a graph is drawn for each item of the percentages of passes in four or more groups, the test constructor can usually judge by eye which items are unsatisfactory (Hawkes and Lindquist, 1937; Thomson, 1942), since their slope will be too gentle or will show reversals. But the percentages, particularly for extreme items of low balance, are likely to be somewhat unreliable, even with 100 cases in each group, hence it is desirable to compute some numerical index of correlation with the criterion, and its statistical significance. Henry (cf. Long, 1935) describes a rough index for assessing the consistency of p 's in successive decile groups.

2. *Application of the Constant Psycho-physical Method.* Burt (1921) has pointed out the similarity of item-analysis data to the results of psycho-physical experiments : a threshold (corresponding to item-difficulty) and a standard deviation (corresponding

to discrimination) can be derived, as in the Constant Method. Ferguson (1942) and Lawley (1943) develop the formulæ for this technique.

3. *Tetrachoric Correlation.* The present writer usually determines two values of r_t , by contrasting p_U in the top tertile with p_L in the two bottom tertiles, and p_U in the top two with p_L in the bottom tertile. These percentages can be obtained by slide-rule, and Burt provides an abac (Vernon, 1946) connecting p_U , p_L and r_t . Averaging the two coefficients yields a single discrimination index which has been called the 'double tetrachoric' coefficient, and the difference between them indicates whether the item is functioning more effectively in the upper or lower part of the scale.

All indices become less reliable as balance decreases, but r_t breaks down altogether if $p_U = 1.0$ or $p_L = 0$. The writer's practice is then to use only one coefficient; thus if percentages of passes in the three tertiles are 100, 96 and 90, he takes 98 as p_U and 90 as p_L .

4. *Chi-squared, phi, and Pearson (product moment) r.* Chi-squared, or equally analysis of variance, can be used for testing the significance of differences in p between successive groups. Conversion to ϕ or r , respectively, yields discrimination indices. These methods can be employed with a categorical criterion, and have often been applied in America to top and bottom 27 per cent. groups. When only two groups are contrasted:

$$\chi^2 = \frac{N(p_U - p_L)^2}{4pq}; \quad \phi = \frac{p_U - p_L}{2\sqrt{pq}}, \text{ where } p = \frac{p_U + p_L}{2}.$$

Guilford (1941) provides an abac for reading off ϕ , and Jurgenson (1947) tabulates ϕ for each percentage value of p_U and p_L . If both distributions are dichotomous, product-moment r is identical with ϕ . Burt (1921) proposed using Yule's coefficient of colligation, and provided an abac for reading it.

5. *Percent. Differences.* Symonds, Holzinger and others (cf. Long, 1935) have proposed indices which effectively amount to $p_U - p_L$.

6. *P.E. of Percent. Differences.* Votaw (1933) and Arnold (1935) give formulæ and abacs or nomographs for reading off the probable error of $p_U - p_L$ in the top and bottom 27 per cents.

7. *Sigma Differences.* Lawshe (1942) and Lawshe and Mayer (1947) employ a nomograph giving distances in σ units between the ordinates cutting off p_U and p_L .

8. *Biserial r Based on Percent Differences.* Flanagan (1939) has converted p_U, p_L differences into approximate biserial correlations, and provides charts for reading off these coefficients.

9. *z-Method.* Davis (1946a) points out that equal increments in r_{bis} do not represent equal increments in discriminating power. He therefore transformed r to Fisher's z , and converted the z 's to a scale with a mean of 50 and a standard deviation of 21. He provides a table from which may be read off the difficulty (also expressed in σ units) and discrimination corresponding to all values of p_U and p_L from 1 to 99.

IV. DISTRIBUTION METHODS

10. *Overlapping Methods.* Guilford (1936a) outlines Vincent's (1924), Long's and other methods derived from the proportion of W 's whose criterion scores exceed the median score of the R 's. These do not appear to have any advocates nowadays.

11. *d-Method.* The simple difference between mean criterion scores of R 's and W 's was suggested by Swineford (1936), and is the chief method used by Anstey in this country. By dividing by σ (cf. Zubin, 1934), the indices for different tests become comparable, and usually range from zero, or negative values, to about 2.0.

12. $r_{bis} = \frac{d}{\sigma} \frac{pq}{z}$. This appears to be the most popular technique among test constructors in the American Forces (cf. Gulliksen, 1944). Its main disadvantage is that exaggerated coefficients, often exceeding 1.0, may result if the distribution of criterion scores in the whole population is not strictly normal, though this can often be avoided by redistributing arbitrarily into a normal shape. Richardson and Stalnaker also criticise the assumption that passing or failing an item represents a normally distributed, continuous variable. For these reasons many statisticians prefer:

13. *Correlation Ratio, Point Biserial, Critical Ratio or F-ratio.* When two groups are contrasted, these are all equivalent, though the first two yield discrimination coefficients, the last two indices of significance.

$$\text{Point Biserial} = \eta = \frac{d\sqrt{pq}}{\sigma}; t = \eta\sqrt{N-1}; F = t^2.$$

If there are several groups, as in tests where the alternative responses to an item receive different amounts of credit, η or F -ratio are particularly appropriate (cf. Lev, 1938).

14. *McCall, Long, Bliss Method.* This too was developed for dealing with complex items. For items which are only marked R or W (or F), the formula is $\frac{d}{\sigma} \frac{pq}{z}$.

V. OTHER METHODS

Unig. Walker (1931-40) defines a test as *unig* when "each score x is composed of correct answers to the x easiest questions, and therefore of correct answers to no other questions." All tests however show some departure from this composition, and the amount of randomness or higgledy-piggledyness of response he calls *hig*. Obviously *unig* is closely linked with internal consistency, and a test of high consistency could be constructed by rejecting items which largely contribute to its *hig*. Though Walker himself has not attempted this, it underlies the recently popularized technique of:

Scalogram Analysis. Guttman (1947) provides methods for ensuring the 'unidimensionality' of a scale or test. Its items are said to be 'reproducible' if, from each testee's rank position on the total score, his response to each item can be reproduced with 90 per cent. accuracy. Though the technique is valuable for short attitude scales, it is extremely tedious for long tests, and, in effect, merely provides a guarantee of high internal consistency. Eysenck and Crown (1948) point out, indeed, that for tests containing 24 or more items, 90 per cent. reproducibility is unnecessarily and almost impossibly stringent.

Factor Analysis. Whether all test items are measuring a single trait or ability, or a mixture, can be explored by factorial analysis of the item inter-correlations (Burt, 1939, Eysenck, 1948). This technique too is scarcely practicable when there are more than 20 to 30 items, and may be superfluous if the item consistency coefficients (corresponding to their general factor saturations) are high. But it is useful for a short test or scale where lowish consistencies suggest that items are measuring two or more distinguishable traits, and for analysing sub-tests—i.e. groups of items. Thus it has chiefly been applied to the Stanford-Binet and Terman-Merrill scales, by Burt (1939), Burt and John (1942), McNemar (1942), Banks and others.

It should be noted that none of the above techniques is applicable to analyses of item validity.

We may now compare the Grouping and Distribution techniques under several headings.

VI. DISCUSSION

(a) *Speed and Reliability.* The tabulations required for Type *A* (Grouping) methods are obviously less onerous than those for *B* methods; and the calculation of percentages and the operations involved in reading off indices from abacs or tables are more rapid than the calculation of means and biserials. Both types are susceptible to various short-cuts. If, for example, the test constructor is interested only in *R* and *W* responses, then, using *A* methods, he can tabulate *W*'s for his upper groups and *R*'s among his lower groups. Equally, with *B* methods, he can tabulate the criterion scores of passes on the difficult items and of fails on the easy items. Biserial or Point Biserial can be derived from $M_r - M_n$ or $M_n - M_w$, just as easily as from $M_r - M_w$ (Dunlop, 1936; Chapanis, 1941).

The *A* methods, however, make less complete use of the available data and for that reason must inevitably be less reliable than *B* ones. Davis (1946*b*) carried out parallel analyses by Method 9 in two groups of 370 testees, and found that, while the difficulty indices from the two groups were closely similar, discrimination indices were more discrepant. He recommends a minimum population of 800 for item analysis—a number which would impose a serious strain on test constructors when machine-scoring is not available. An experiment on the relative superiority of *B* methods is described below.

(b) *Multiple Criteria.* When items are being compared with two or more criteria, say an internal and an external one, the amount of tabulation needed for Type *B* indices is doubled. But with a Type *A* method such as No. 3, all that is needed is to sort the answer sheets into nine piles instead of three, and the tabulation is unaltered. The first pile will contain testees in the top tertile for both criteria, the second pile top on one criterion and middle on the other, and so on. Type *A* methods involving top and bottom 27 per cent. groups will involve extra tabulating depending on the smallness of the correlation between the criteria.

(c) *Detailed Analysis.* When the surplus of items is not very large, and it is desired to revise and improve weak ones, a more detailed analysis than that based on *R* and *W* (or *R* and *F*) is worth-while. In multiple-choice tests the responses of testees giving each wrong answer should be tabulated, in order to see that no one alternative is over-popular. For as Horst (1933) has shown, item-reliability is increased when all the wrong alternatives are equally popular. One alternative, again, may be poorer than others in the sense that it is chosen by many of the better testees. In an inventive-response test there is often some doubt as to whether to allow certain alternative answers as correct, and this can be decided by tabulating the criterion scores of testees who give such alternatives. The most straightforward and reliable technique for this kind of item analysis is the *d*-method, and there is little difficulty in turning the means for testees who give various alternatives into biserials. There is no reason, however, why it should not be tackled by Type *A* methods. In fact Anstey has found that, of the items shown by the *d*-method to be unsatisfactory or to require revision, four-fifths can be picked out by using ϕ coefficients between top and bottom tertile groups, provided that allowance is made for the influence on ϕ of imbalance (cf. (g) below).

(d) *Guessing Correction.* In multiple-choice items the percentages of passes are inflated by guessing. With Type *A* indices this can readily be corrected, and Guilford (1936*b*) recommends the formula:

$$\text{Corrected } p = \frac{R - \frac{W}{n-1}}{N}.$$

Davis (1946a) prefers to disregard U responses, hence his version is :

$$\text{Corrected } p = \frac{R - \frac{F}{n-1}}{T}.$$

Whether the same adjustment should be made to p in the biserial and point biserial (Type B) methods does not seem to have been investigated. With the d -method, the problem does not arise.

(e) *Spurious Internal Consistency Correlations.* In determining the consistency index of an item, that item is itself included in the criterion. Zubin (1934) provides formulæ for removing this spurious element from Type B indices, but there is no easy way of allowing for it in Type A methods. On the whole this defect seems quite trivial since, with a test of say 50 or more items the effect of omitting or including one item is very small. Moreover it applies to all items, and as we are interested mainly in the relative consistency of different items in a single test, it merely boosts all their indices to much the same extent.

(f) *Effects of Insufficient Time.* It is often impossible to apply a test with sufficient time for all testees to try all items. When the poorer subjects fail to reach the more difficult items towards the end of the test, all indices tend to be seriously distorted. If R 's and W 's alone are considered, the indices will be boosted, since all the poorer subjects will be credited with failing such items though some of them might actually have passed a few. On the other hand if R 's and E 's or F 's alone are used, indices will be unduly lowered, for the E 's on difficult items will be made up only of the better subjects who reached these items. It was pointed out above that it is often impossible to decide whether an omitted item is O or U , hence F is a rather 'chancy' quantity upon which to base any index. Moreover the use of E or F instead of W raises additional problems. T (that is $R + F$) will differ for each of the later items, hence indices based on probability such as chi-squared or t , which depend on numbers of cases, will no longer be comparable throughout. Indices involving σ , such as biserial or point biserial, will take much longer to calculate, since σ will differ for each item. And Type A indices may break down since the bottom 27 per cent. or 33 per cent. on the criterion may consist almost wholly of U 's and contain few if any E 's. The d -index does not, like r_{bis} , involve extra calculation, since there is no need to divide it by σ . But for that very reason it will, if based on T , drop all the more rapidly as U increases.

The moral is, presumably, that ample time should always be allowed in experimental trials of a test. However, provided items are well graded in order of difficulty and the time is long enough for at least half the testees to finish, it is unlikely that serious inaccuracies will arise from analysing R 's and W 's only, and treating all O 's and U 's as E 's. In the case of a speed test, where the score definitely depends on the number of (what are usually) easy items accomplished within a given limit, item analysis is scarcely necessary.

In intermediate cases, that is in power tests where only a small proportion have been able to finish, the best plan would seem to be to use the d -method, but to tabulate separately the criterion scores of R 's, O 's, E 's and U 's. A fairly safe indication of item discrimination can then be obtained for later items by considering $M_r - M_w$ and $M_r - M_f$, and possibly averaging them or combining them in such a way that the indices neither rise nor fall markedly as U increases. If r_{bis} is preferred to d the two d 's can be averaged, and p obtained from $\frac{R}{2} \left(\frac{1}{N} + \frac{1}{T} \right)$.

(g) *Effects of Balance.* The only indices not affected by balance are r_t , r_{bis} and Sigma Differences (Nos. 3, 7, 8, 9, 12). Most methods based on $p_U - p_L$, together

with φ , η and Point Biserial favour middling items. Since $\eta = r_{bis} \frac{z}{\sqrt{pq}}$ and $d = r_{bis} \frac{\sigma z}{\sqrt{pq}}$, it can readily be deduced that the d -method favours extreme items. If, for example, we take items of different balance, for all of which the r_{bis} index = 0.5, we get the following η 's and d 's.

p or q	η	$\frac{d}{\sigma}$
.50	.399	.798
.40	.394	.805
.30	.379	.828
.20	.350	.875
.10	.293	.975
.05	.237	1.086

Plotting $\frac{d}{\sigma}$ against η , the points fall on a straight line, showing that the effects of imbalance are equal and opposite.

Several investigators have reported correlations with balance. Thus in a study by Henry (Long, 1935), the McCall, Long, Bliss index correlated + 0.568, Upper and Lower Thirds ($p_U - p_L$) + 0.558, d - 0.385, and r_{bis} + 0.105. Swineford (1936a, 1936b) similarly obtained coefficients of about .5 for indices based on $p_U - p_L$; Pearson r gave + .271, d - .183. Little importance, however, should be attached to the absolute sizes of such coefficients, for they depend (as will be shown below) on the range of difficulty and of discrimination in the particular set of items analysed. Moreover it is often easier to construct good middling items than very easy or very difficult ones; and we have already seen that indices for difficult items at the end of a test are distorted if the time limit is inadequate.

Long (1935) and Swineford (1936a) point out that indices which favour middling items are advantageous in tests for homogeneous groups where discrimination is chiefly required near the median. In effect such indices reflect, not merely the goodness of the item but also the numbers of individuals between whom the item discriminates. Richardson (1936) and Gulliksen (1945) have also proved that the variance and discriminating power of a test increase the higher the average inter-correlation of the items and the more nearly the item-difficulties cluster around .5. When however a test is designed to discriminate over a wide range of ability in heterogeneous groups, difficulty too should range more widely and discrimination indices should be independent of balance. Otherwise it will be possible to compare discrimination indices and to select the best items only at each particular difficulty level. (This can often be done quite effectively by plotting the discrimination indices of the items against their difficulty.) There are psychological advantages also in including items with a fairly wide spread of difficulty. Both the poorest and best testees may be put off if the difficulty is close to .5 throughout.

In no practical situation does it appear desirable to weight the value of the extreme items, as the d -method does. But it is usually possible to compensate for this tendency also by plotting discrimination against difficulty.

(h) *Selection of items for Maximum Validity.* In vocational and educational selection it is generally realized that aptitude tests should have high correlations with subsequent proficiency but low correlations with one another. The component items of a test should be regarded in the same way as the component tests of a selection battery. Merely choosing the most valid items (against an external criterion) does not necessarily produce the most valid test (cf. Smith, 1934; Long, 1935); though as an experiment by Kroll (1940) has shown this does tend to occur when the range of item-validities is a wide one. It is seldom practicable however to compute all

the item inter-correlations and to apply multiple correlation technique for picking the best combination of items. Most test constructors seem piously to hope that the inclusion of a wide and representative range of content in their items will sufficiently reduce their item inter-correlations. Yet often, as we have seen, they use an internal instead of an external criterion, or else employ Scalogram Analysis, and these naturally increase, not reduce, item inter-correlations. Thomson (1942) suggests that, in constructing intelligence tests, internal and external criteria yield much the same results. But this would not be true of all types of test, and in the experiment with an intelligence test reported below, the correlation between consistency and validity indices (r_i 's) was only + 0.533. If, for example, the best 60 had been selected from the 90 analysed items, a quarter of those indicated by the internal criterion would have been excluded by the external criterion, and vice versa. The two criteria agreed only on 45 of the best items.

Toops's (1941) *L*-method and Horst's (1934) method of successive residuals depend on building up successive composites of the most highly valid items, but though they do not require all the item inter-correlations they are extremely elaborate. A simpler technique introduced by Horst (1936) and frequently adopted by American psychologists in the Services depends simply on contrasting r_{IC} (the validity) with r_{IT} (the consistency). Richardson and Adkins (1938) similarly advocate the use of the formula :

$$\frac{r_{IC} - r_{IT}r_{TC}}{(r_{TC} - r_{IC}r_{IT})\sigma_I}$$

In other words, consistency is used to indicate the badness rather than the goodness of an item.

Much depends, of course, on the type of test. An internal criterion is legitimate if the validity can be taken for granted, and this is commonly the case with educational achievement tests which include a fair sampling of the whole field of knowledge at which the test is aimed. The same might apply to many tests of social attitudes. But it is unsatisfactory with tests of mechanical or other aptitudes, although admittedly it may be exceedingly difficult to attain an external criterion which is not unreliable or biased, for a large enough population of testees (cf. Vernon and Parry, 1948). If the validity of the test is low, using an internal criterion to select the most consistent items is merely likely to make the revised test a more reliable measure of something invalid. However this is a problem which has little bearing on the choice of discrimination indices. Horst's and Richardson's methods have chiefly been applied to biserials, but could presumably also be used with tetrachorics, ϕ coefficients or point biserials.

VII. EXPERIMENTAL STUDIES

Numerous investigations were carried out in the 1930's into the correlations between different indices applied to the same item-data, and the reliability of tests selected by such indices (Barthelmeß, 1931 ; Lentz and Hirshstein, 1932 ; Lindquist and Cook, 1933 ; Henry, Long, 1935 ; Swineford, 1936*a, b* ; Pintner and Forlano, 1937). Long points out that none of these can be conclusive since the most suitable method varies with the type of test required (e.g., a high or low concentration of middling items), with the level of ability of the population and with the level of difficulty of the items. Because one method gives the best results in one research it does not follow that it would do so if easier items of the same kind were tried out on a less able population. Long indeed concludes that most indices overlap so closely (apart from the balance factor) that ease of tabulation and computation should be a main consideration in choosing between them. The present writer's excuse for

adding to these studies is to see how the double tetrachoric method, which has not previously been described, compares with the standard methods. Further objects were :

(a) to find the relative times taken over tabulations and computations for *A* and *B* type indices ;

(b) to verify the dependence of certain methods on balance ;

(c) to investigate the reliability of indices based on the two types of method.

Analyses were made of three tests devised for adults of average or rather above average ability, who ranged however from very low to very high levels.

(i) A 33 multiple-choice item vocabulary and reading test applied to 134 Ordinary Seamen, without time limit.

(ii) Three versions of Number and Letter Series (inventive-response) tests, each containing 30 items. These were applied, with 15 minute limits, to parallel groups of 373 naval, army, R.A.F., and W.A.A.F. recruits.

(iii) A 182 multiple-choice item general intelligence test given, without time limit, to 200 Civil Service clerical assistants. This experiment was undertaken by the Research Unit of the Civil Service Commission.

An internal criterion was used in all three groups, but in the second one an external criterion was also employed, namely the score on the standard battery of intelligence tests used by each Service (*T2*, Total *S.G.* and Total *G.V.K.* respectively).

(a) *Tabulation Time.* The tabulation of Group (i) responses occupied one statistician, Miss M. S. Stevenson, $2\frac{3}{4}$ hours for double tetrachorics and $6\frac{1}{2}$ hours for the *d*-method or r_{bis} . The second operation seemed rather less tedious to her in spite of its greater length. Turning the tabulated data into percentages and r_t 's, and calculating the means for the *d*-method, each took about $1\frac{1}{2}$ hours.

In Group (ii) all tabulations were done by Hollerith (Admiralty Statistics Branch). Though times are not available it was, if anything, simpler to total the criterion scores of the passes on each item than to sort the testees, according to criterion scores, into three groups and to count the passes in each group. Biserial *r* was also very simply obtained from such totals by the formula :

$$\frac{pX - X_r}{N\sigma_z}$$

where *X* = the total criterion scores of the whole group, *X_r* the total scores of testees getting the item right.

In Group (iii) several statistical assistants were involved. The tabulations for r_t occupied 12 man-hours and the calculations another 7 hours.* Tabulations and calculations for the full *d*-method (where each answer to each alternative is listed separately) totalled 38 man-hours. Thus though this also took twice as long as r_t , more detailed information than merely $M_r - M_w$ was obtained. A third method was tried, namely ϕ coefficients based on the top and bottom thirds of 300 cases, i.e., a total of 200. Tabulation again took 12 hours, but calculations only 3 hours ; and when each answer was treated separately the tabulation time rose to 19 hours. We may conclude then that even when detailed analysis is involved, Type A methods require only about half the time of Type B.

(b) *Inter-correlations and Correlations with Balance.* In Group (i) rank-order correlations were calculated between six indices of item consistency, and between each index and balance. These are shown in Table I. The coefficients are very high,

* In the writer's opinion the latter figure is unnecessarily high. With practice the determination of percentages by slide-rule, looking up r_t 's in an abac, and averaging them takes little if any longer than finding ϕ 's (with interpolations where needed) from Jurgenson's tables.

TABLE I. ITEM-CONSISTENCY INDEX INTER-CORRELATIONS

	$p_U - p_L$	ϕ	η	r_t	r_{bis}	d	Balance
$p_U - p_L$		.91	.90	.76	.70	.50	+.47
ϕ or χ^2			.94	.89	.81	.63	+.30
η				.88	.90	.73	+.18
r_t					.93	.86	-.04
r_{bis}						.95	-.16
d							-.36

except those between indices which are positively and negatively affected by balance ; and r_t shows at least as good agreement as r_{bis} with other indices.

With only 33 items we cannot expect reliable correlations with balance, but they do follow the expected order. Percent differences and ϕ favour middling items, and d extreme ones. η also tends to favour middling though only to an insignificant extent, and r_{bis} appears to correlate slightly with imbalance. Conceivably this particular test contains better extreme items than middling ones. This would explain why the coefficient for r_{bis} is slightly negative and why that for η is not more strongly positive.

In Group (iii) the ϕ coefficients showed a high correlation with balance, but by graphing them against difficulty it was possible to pick out the weakest items, or those which required revision. Indeed a detailed analysis by the ϕ method yielded much the same agreement with the d -method as did the r_t analysis.

In Group (ii) the following product-moment correlations were obtained between balance and consistency and validity indices.

	ϕ	η	r_t	r_{bis}	d
Internal Criterion ..	+.837	+.737	+.217	+.102	-.492
External Criterion ..	+.711	+.554	+.145	+.176	-.133

It may be seen that they are quite similar, though the external ones are more attenuated. The coefficients tend to be much larger than those reported for Group (i). The probable reason is that these test items did not vary widely in discrimination, hence balance exerted a relatively large effect. In Group (i), on the other hand, variations in consistency were large relative to variations in difficulty.

Here the coefficients for r_t and r_{bis} are all positive, though small ; in Group (i) they were both negative. This suggests that with a more representative selection of items, they should both be zero. Inspection of the scatters showed that the positive correlations were due wholly to the easy items being rather less discriminative than the middling ones. The difficult items obtained quite as high consistency or validity indices as the middling ones.

As in Group (i) the coefficients for ϕ are even more strongly correlated with balance than those for η . The coefficients for d are negative, though not, as predicted, equal to the η coefficients. This is probably attributable again to the weakness of the easy items. If we compensate by making the mean of r_t and r_{bis} coefficients

zero, and subtract the difference throughout, we get :

	ϕ	η	$r_t + r_{bis}$	d
Internal Criterion ..	+·677	+·577	·000	—·652
External Criterion ..	+·550	+·393	·000	—·294

Apart from signs, η and d no longer differ significantly, with a population of 90 items.

(c) *Reliability*. The sampling errors of r_{bis} indices are the same as those of t , and are given by : $S.E.$ of zero $r_{bis} = \frac{\sqrt{pq}}{z\sqrt{N}}$. But the $S.E.$ of a zero tetrachoric correlation, when one variable is divided at q , the other at the 33rd or 67th percentile, is $\frac{\sqrt{.333 \times .667}}{.3636} \times \frac{\sqrt{pq}}{z\sqrt{N}}$. This is 1·297 times greater than the $S.E.$ of r_{bis} at all difficulty levels. Thus if an item obtains a significant d or r_{bis} index with 100 cases, $100 \times 1\cdot297^2 = 168$ cases would be needed for the same item to give a significant r_t . However the decrease in reliability for the average of two tetrachorics is not so large. When both splits are at the 33rd percentile it has been found by the writer (Vernon, 1948) that the $S.E.$ of a double tetrachoric is 1·323 times smaller than that of the single coefficients on which it is based.

A study of the relative reliability of r_{bis} and double r_t was carried out by dealing a set of 90 playing cards 50 times. The cards were chosen to yield an approximately normal distribution, namely :

Ace and Queen ..	..	1
2 and Knave ..	..	3
3 and 10 ..	..	5
4 and 9 ..	..	9
5 and 8 ..	..	12
6 and 7 ..	..	15
		—
		90

By grouping Ace-5, 6-7, and 8-Queen together, three sets of 30 could be obtained for yielding double tetrachorics. Only one level of item-difficulty was studied, $p = .20$. Thus the distribution was recorded for the first 18 cards after each reshuffling. From the difference between the means of such distributions and the mean for all 90 cards (namely 6·5), a series of fifty biserials was obtained ; and from the numbers of Ace-5's, 6-7's and 8-Queen's, one hundred single and fifty double tetrachorics were calculated.

The $S.E.$'s predicted by formula are $\frac{1\cdot4287}{\sqrt{N}}$ for r_{bis} and $\frac{1\cdot8524}{\sqrt{N}}$ for single r_t . The obtained $S.E.$'s were 1·5159 and 1·9752 respectively. These are 1·062 and 1·066 times too large, but the agreement with the formulae is quite good considering the small number of cases and the difficulty of really thorough shuffling.

The obtained $S.E.$ for double tetrachorics was $\frac{1\cdot6782}{\sqrt{N}}$. While considerably smaller than the figure for single r_t 's it is still somewhat larger than that for r_{bis} , namely 1·1071 times. Thus for an item whose r_{bis} index is significant with 100 cases, double r_t would be significant with 123 instead of 168 cases.

The same data were used for exploring the significance of indices such as Nos. 4-8, based only on the top and bottom 33 per cent., disregarding the middles. (In practice, of course, 27 per cent. tails are used, but the results should be fairly similar.) The mean value of χ^2 for the 50 deals was 1·1712, with $N = 60$. Thus N would need to

be 339 for χ^2 to reach the .01 level of significance, and the total population from which the top and bottom tertiles were drawn would be 509. Now the (arithmetical) mean value of d was .550, which yields a mean t of .900, with $N = 90$. Thus for t or r_{bis} to reach .01 significance, only 258 cases would be needed. In other words, if a Type B index was significant with 100 cases, a Type A index (other than r_t) would require 131 cases drawn from a population of 196. Thus χ^2 or ϕ and double r_t have, as one would expect, much the same reliability, and although Type B indices are more reliable the difference is not large. But when Type A indices are based only on the tails of the distribution, and the middles are not tabulated, the total population requires to be about twice as big as for Type B indices. It follows therefore that Type B indices should certainly be used when the population available for analysis is small, say 200 or less. When it is somewhat larger r_t should be adequate, but if account is taken only of the tails much larger populations are desirable.

VIII. SUMMARY AND CONCLUSIONS

1. Current methods of analysing the validity or consistency of test items are listed. They fall under two main headings : (A) Grouping methods, where the item responses of several groups of testees having different criterion scores are tabulated ; (B) Distribution methods, where the distributions of criterion scores are tabulated for testees giving each response to an item. Unig and scalogram analyses of item consistency and factor analysis are considered to have much more limited applicability.

2. A new method under Type A, advocated by the writer, is the 'double tetrachoric,' that is the mean of tetrachoric correlations between passing the item and the criterion, when the latter is split at approximately the 67th and 33rd percentiles. Indices obtained by this method are shown to correlate highly with other kinds of indices.

3. Type A indices based on proportions of item passes in the top and bottom 27 per cent. groups, such as Flanagan's and Davis's, also r_t , and the r_{bis} index in Type B, are independent of the difficulty (balance) of the items. But Type A indices such as χ^2 and ϕ , and Type B indices such as Point Biserial or eta, strongly favour items of middling difficulty. They are therefore less useful for most test construction purposes, in spite of their greater popularity among statisticians.

4. The simplest Type B method, namely the difference in means between testees who pass and fail an item (the 'd-method') favours extreme items as much as eta favours middling ones, though this tendency is often obscured by the relative weakness of the very easy or very difficult items.

5. Experimental trials show that the tabulations and calculations for Type A indices take only about half as long as those for Type B. The r_t index is particularly saving of time when items are analysed against more than one criterion. But the d-method or r_{bis} is more appropriate when all the responses given to an item are analysed instead of merely the rights and wrongs.

6. If a test has been tried out with insufficient time for all testees to finish, all indices are distorted, but it is probably easier to compensate when Type B indices are employed.

7. With selective-response tests, guessing corrections are more readily applicable to Type A methods. Spurious internal consistency, due to the inclusion of the item analysed in the criterion, can be more readily corrected with Type B methods but such correction is usually unnecessary. The problem of choosing the most valid set of items for a revised test, as distinct from the most consistent, is not affected by the type of index used.

8. The r_t index is somewhat less reliable than r_{bis} or d , and other Type *A* indices considerably less so. It is shown that when the r_{bis} index is statistically significant with 100 cases, 123 cases would be needed for r_t to be equally significant. The total population for other Type *A* methods would have to be approximately 200, although only the responses of the tails are tabulated.

9. Our main conclusions, therefore, are that Type *B* methods should be used when numbers of testees are unavoidably small, when detailed analysis of responses is desired, and when time limits have been markedly insufficient. The d -method is the simplest of these if allowance can be made for its favourableness to items of high and low difficulty, but r_{bis} gives indices independent of 'balance,' and Point Biserial is preferable when items of middling difficulty are to be favoured.

Type *A* methods are much quicker; and of these double tetrachoric requires the smallest number of cases. Hence they are advocated when there is a surplus of items and an analysis of Right vs. Wrong responses will suffice for selecting the most discriminative ones.

REFERENCES

1. Arnold, J. N. 'Nomogram for determining the validity of test items.' *J. Educ. Psychol.*, XXVI, 1935, 151-153.
2. Barthelme, H. M. 'The validity of intelligence test elements.' *Teach. Coll. Contr. Educ.*, 1931, No. 505.
3. Burt, C. L. *Mental and Scholastic Tests*. London: King, 1921, 432.
4. Burt, C. L. 'The latest revision of the Binet intelligence tests.' *Eugen. Rev.*, XXX, 1939, 255-260.
5. Burt, C. L., and John, E. 'A factorial analysis of the Binet-Simon tests.' *Brit. J. Educ. Psychol.*, XII, 1942, 117-127.
6. Chapanis, A. 'Notes on the rapid calculation of item validities.' *J. Educ. Psychol.*, XXXII, 1941, 297-304.
7. Davis, F. B. 'Item-analysis data; their computation, interpretation, and use in test construction.' *Harvard Educ. Papers*, No. 2. 1946a, 42.
8. Davis, F. B. 'Notes on test construction; the reliability of item analysis data.' *J. Educ. Psychol.*, XXXVII, 1946b, 385-90.
9. Dunlop, J. W. 'Note on computation of biserial correlations in item evaluation.' *Psychometrika*, I, 1936, 51-60.
10. Eysenck, H. J., and Crown, S. 'An experimental study in opinion-attitude methodology.' *Int. J. Opin. Attitude Res.*, 1948. To be published.
11. Ferguson, G. A. 'Item selection by the constant process.' *Psychometrika*, VII, 1942, 19-29.
12. Flanagan, J. C. 'General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution.' *J. Educ. Psychol.*, XXX, 1939, 674-680.
13. Guilford, J. P. *Psychometric Methods*. New York: McGraw, Hill, 1936a, 566.
14. Guilford, J. P. 'The determination of item difficulty when chance success is a factor.' *Psychometrika*, I, 1936b, 259-264.
15. Guilford, J. P. 'The phi coefficient and chi square as indices of item validity.' *Psychometrika*, VI, 1941, 11-19.
16. Gulliksen, H. 'Selection of test items by correlation with an external criterion, as applied to a mechanical comprehension test.' 1944; *O.S.R.D., Publ. Bd. No. 13319*. Washington, D.C.: U.S. Dept. Commerce, 1946, 11.
17. Gulliksen, H. 'The relation of item difficulty and inter-item correlation to test variance and reliability.' *Psychometrika*, X, 1945, 79-91.
18. Guttman, L. 'The Cornell technique for scale and intensity analysis.' *Educ. Psychol. Measmt.*, VII, 1947, 247-280.
19. Hawkes, H. E., Lindquist, E. F., and Mann, C. R. *The Construction and Use of Achievement Examinations*. London: Harrap, 1937, 497.
20. Horst, P. 'The difficulty of a multiple choice test item.' *J. Educ. Psychol.*, XXIV, 1933, 229-232.
21. Horst, P. 'Item analysis by the method of successive residuals.' *J. Exper. Educ.*, II, 1934, 254-263.
22. Horst, P. 'Item selection by means of a maximising function.' *Psychometrika*, I, 1936, 229-244.

Indices of Item Consistency and Validity

23. Jurgenson, C. E. 'Table for determining phi coefficients.' *Psychometrika*, XII, 1947, 17-30.
24. Kelley, T. L. 'The selection of upper and lower groups for the validation of test items.' *J. Educ. Psychol.*, XXX, 1939, 17-24.
25. Kroll, A. 'Item validity as a factor in test validity.' *J. Educ. Psychol.*, XXXI, 1940, 425-436.
26. Lawley, D. N. 'On problems connected with item selection and test construction.' *Proc. Roy. Soc., Edinburgh, Section A*, LXI, 1943, 273-287.
27. Lawshe, C. H. 'A nomograph for estimating the validity of test items.' *J. Appl. Psychol.*, XXVI, 1942, 846-849.
28. Lawshe, C. H., and Mayer, J. S. 'Studies in item analysis; the effect of two methods of item validation on test reliability.' *J. Appl. Psychol.*, XXXI, 1947, 271-277.
29. Lentz, T. F., Hirshstein, B., and Finch, J. H. 'Evaluation of methods of evaluating test items.' *J. Educ. Psychol.*, XXIII, 1932, 344-350.
30. Lev, J. 'Evaluation of test items by the method of analysis of variance.' *J. Educ. Psychol.*, XXIX, 1938, 623-630.
31. Lindquist, E. F., and Cook, W. W. 'Experimental procedures in test evaluation.' *J. Exper. Educ.*, I, 1933, 163-185.
32. Long, J. A., and Sandiford, P. 'The validation of test items.' University of Toronto, *Dep. of Educ. Res., Bull.* No. 3, 1935, 126.
33. McNemar, Q. *The Revision of the Stanford-Binet Scale*. Boston: Houghton Mifflin, 1942, 189.
34. Pintner, R., and Forlano, G. 'A comparison of methods of item selection for a personality test.' *J. Appl. Psychol.*, XXI, 1937, 643-652.
35. Richardson, M. W. 'The relation of difficulty to the differential value of a test.' *Psychometrika*, I, 1936, 33-49.
36. Richardson, M. W., and Adkins, D. C. 'A rapid method of selecting test items.' *J. Educ. Psychol.*, XXIX, 1938, 547-552.
37. Richardson, M. W., and Stalnaker, J. M. 'A note on the use of biserial r in test research.' *J. Gen. Psychol.*, VIII, 1933, 463-465.
38. Smith, M. 'The relationship between item validity and test validity.' *Teach. Coll. Contr. Educ.*, No. 621, 1934.
39. Swineford, F. 'Validity of test items.' *J. Educ. Psychol.*, XXVII, 1936a, 68-78.
40. Swineford, F. 'Biserial r versus Pearson r as measures of test-item validity.' *J. Educ. Psychol.*, XXVII, 1936b, 471-472.
41. Thomson, G. H. 'Following up individual items in a group intelligence test.' *Brit. J. Psychol.*, XXXII, 1942, 310-317.
42. Toops, H. A. 'The L -method.' *Psychometrika*, VI, 1941, 249-266.
43. Vernon, P. E. 'Statistical methods in the selection of Navy and Army personnel.' *J. Roy. Statist. Soc.*, VIII, 1946, 139-153.
44. Vernon, P. E. 'Classifying high-grade occupational interests.' *J. Abn. Soc. Psychol.*, 1948. To be published.
45. Vernon, P. E., and Parry, J. B. *Personnel Selection in the British Forces*. London: University of London Press, 1948. To be published.
46. Vincent, L. 'A study of intelligence test elements.' *Teach. Coll. Contr. Educ.*, No. 152, 1924.
47. Votaw, D. F. 'Graphical determination of probable error in validation of test items.' *J. Educ. Psychol.*, XXIV, 1933, 182-186.
48. Walker, D. A. 'Answer-pattern and score-scatter in tests and examinations.' *Brit. J. Psychol.*, XXII, 1931, 73-86; XXVI, 1936, 301-308; XXX, 1940, 248-260.
49. Zubin, J. 'The method of internal consistency for selecting test items.' *J. Educ. Psychol.*, XXV, 1934, 345-356.

THE *d* METHOD OF ITEM-ANALYSIS

By E. ANSTEY

Civil Service Commission Research Unit

I. *Object and Limitations of Item-Analysis.* II. *Notation.* III. *The d Method.* IV. *Technique for Examining Doubtful Items.* V. *Fifteen Types of Fault in Item Construction.* VI. *Summary.*

I. OBJECT AND LIMITATIONS OF ITEM-ANALYSIS

The principal object of item analysis is to obtain objective quantitative evidence about the consistency and validity of test items. This is combined with subjective evidence, and is used to improve a test by

- (a) identifying defects in items, some of which can be corrected,
- (b) selecting the best items, which will be rearranged in an improved order of difficulty in the revised version of the test.

Successful item-analysis has the effect of increasing both the reliability and the validity of a test as a whole. For this purpose it should clearly be carried out before the final version of the test is printed and put into use. Another advantage of item-analysis, however, is that it stimulates ideas on test-construction, and facilitates insight into how subjects have reacted to the test. For this purpose item-analysis of the final version of the test has also some value.

The principle of evaluating the separate items that make up a mental test (in addition to assessing the reliability and validity of the test as a whole) was first proposed and applied by Burt (*Mental and Scholastic Tests*, 1921), in connexion with the original Binet-Simon scale. By making an item-analysis of the initial scale, and rejecting the less satisfactory test-problems, Burt succeeded in increasing the reliability from .68 to .91, the validity from .59 to .76, and the general factor loading from 34 per cent. to 46 per cent. Subsequent investigations by Hull, Uhrbrock and Richardson, and others have reached similar results with other tests.

Useful though item-analysis is, there are four important limitations to what can be achieved by it.

(i) Indispensable requirements for the development of a good test are an accurate assessment of what the test is required to measure, a satisfactory plan for the form and layout of the test, and skill and a fertile imagination in making test items. No technique of item-analysis can turn poor items into good items, or operate efficiently in the absence of a reliable criterion against which to assess the value of the items.

(ii) Item-analysis methods should be used with discrimination and understanding. For example, the usual techniques are liable to give unreliable results if the test is measuring widely different factors, or if the influence of speed on performance is very great. Brogden (*Psychometrika*, 1946) has pointed out that, using total score in the test as criterion for item-analysis, any concentration of test-items at a particular difficulty level tends to favour those items at the expense of the rest. Automatic selection according to the consistency index would thus still further increase the concentration of items at that difficulty level.

(iii) Results obtained by an item-analysis from any one experimental group of subjects would not necessarily hold exactly for another group. Unless the number of cases is very great, little importance can be attached to small differences between the indices obtained for different items.

(iv) Purely quantitative evidence should be used to reinforce and supplement, but not to supplant subjective opinions. If for seemingly logical and fool-proof items the item-analysis results are definitely poor, those items should be rejected; but the converse does not apply. If an item has a logical flaw or other real weakness, it should not be used, however favourable the evidence from item-analysis. To be worth using, an item must be known to work well in practice, but it must also be sound and defensible against theoretical attack.

II. NOTATION

It is convenient to use the standard notation which I suggested while serving in the Army Directorate for the Selection of Personnel and which is used by this Research Unit (R.U.). Capital letters are used for the number of people putting a particular response to an item, small letters for means, proportions, etc.

N is the total Number of cases in the sample being analysed.

T is the number who Tried the item, i.e., who reached the item, whether or not they put an answer down.

U is the number for whom the item was Unattempted, i.e., who did not reach it at all. $T + U = N$.

R is the number who put the Right answer. $R/N = p$.

E is the number who put an answer but an Erroneous one.

O is the number who reached the item (i.e., answered at least one later item) but Omitted to put any answer.

F is the number who tried the item and Failed to put the right answer. $E + O = F$. $R + F = T$.

W is the number who got the item Wrong. $W = E + O + U$. $R + W = N$. $W/N = q$.

m is the mean criterion score of *N*.

m(R) is the mean criterion score of *R*. Similarly for *m(E)*, *m(O)*, *m(W)* etc.

$d = m(R) - m(W)$.

$d(E) = m(R) - m(E)$.

$d(F) = m(R) - m(F)$.

σ is the standard deviation of criterion scores.

D (Burt's "standardized difference between means"¹) = d/σ .

r_b (biserial *r*) = $D \times \frac{pq}{z}$.

III. THE *d* METHOD

(a) *Description of method.* The *d* method of item-analysis requires the calculation of a mean criterion score for each kind of response to an item. It is therefore most obviously applicable when the criterion approximates to a continuous variable: it can however also be used when the criterion consists e.g., of an overall grading on a 5-point scale, or even when it is dichotomous, by giving 'Passes' a score of 1 and 'Fails' a score of 0. In the latter case, given *m(R)* and *m(W)*, tetrachoric *r* can be read straight off an appropriate table. The criterion may be either an external, independent assessment of proficiency, or an internal criterion, i.e., total test score.

¹ Burt, C., 'Item-Analysis in Educational Tests' (unpublished memorandum, 1945).

The characteristic feature of the method is that responses to an item are not merely divided into R and W , but classified into R , E , O and U . Moreover each different Erroneous answer in a choice-response test, and each different E answer put by more than one subject in an inventive-response test is recorded separately. The mean criterion score is calculated for R , E and each different E , O , U , and W , so that a complete picture is available of the responses made to the item by the sample. N should normally be at least 200 to give reliable results. $d = m(R) - m(W)$ is then calculated. D and r_b can also be calculated if required, but normally this is not necessary.

For an ordinary, satisfactory item, the mean criterion scores for R , E , O and U usually rank in that order. If sufficient time has been allowed for the experimental trial, U should be zero except perhaps for the last few items in the test. If U is not zero, $m(U)$ is almost invariably low, as it is the slower and weaker subjects who will have failed to attempt the last items. $m(E)$ is normally slightly higher than $m(O)$, since people who have been able to put some answer to an item are usually slightly better than those who could put no answer, but this is by no means invariably the case, nor does it greatly matter if these means are reversed. The essential point is that $m(R)$ should be substantially higher than the other means. If the order of the means is very irregular, or if $m(R)$ is not substantially higher than the others, all responses to the item must be closely examined in order to try to find the explanation of what has happened.

(b) *Specimen results.* Three 'Abstractions' items from the first draft of the R.U. General Classification Test Part C will be quoted as giving typical results.

Through the co-operation of the War Office D.S.P., the draft test was tried out on a fairly representative group of young soldiers. Using total score in the test as criterion, item-analysis results were as given in Tables I-III. $N = 200$, $m = 9.83$, $\sigma = 6.94$.

TABLE I

Item 3. "Trend, rend, end; scare, care, ***?"
Correct answer 'are,' beheading the previous word.

Item	R	$m(R)$	E	$m(E)$	O	$m(O)$	U	$m(U)$	W	$m(W)$	d	D	r_b
3	82	15.95	87	6.04	30	4.33	1	2.00	118	5.57	10.38	1.500	.933
Analysis of E	{	putting	air	8	9.87		putting	bear	2	6.50			
		dare	5	7.40									
		fare	9	5.78									
		mare	5	7.00									
		rare	6	6.16									
		tare	3	6.00									
		other E	29	4.48									

No fewer than 41 different E answers were put to this item. Those put by two or more subjects are recorded above. The best E answer was "air," which has some justification. Subjects were meant to omit the first letter of "care" and leave the other three letters alone; but these E subjects put "air," presumably in order to preserve the rhyme. Their mean was six below $m(R)$, but this E answer was sufficiently plausible to justify special consideration, as will be explained in paragraph IV(b). Clearly the wording of the item was not well chosen. It would have been better to have used a sequence of words such as drink-rink-ink with a continuity of rhyme and so have avoided any ambiguity. The other E answers have no apparent justification.

All four-letter answers ignored the fact that there were only three asterisks in the clue. "car" omitted the last letter, instead of the first letter of "care." "bend" was a hangover from the first three words of the item. In this instance, despite the diversity of answers, the probable reason behind each *E* response was clear enough.

TABLE II

Item 13.

"Mile, length; acre, ****?"

Correct answer 'area,' since mile is a measure of length and acre a measure of area.

Item	<i>R</i>	<i>m(R)</i>	<i>E</i>	<i>m(E)</i>	<i>O</i>	<i>m(O)</i>	<i>U</i>	<i>m(U)</i>	<i>W</i>	<i>m(W)</i>	<i>d</i>	<i>D</i>	<i>r_b</i>
13	47	19.00	88	7.43	58	6.91	7	2.57	153	7.01	11.99	1.728	1.011*
<i>Analysis of E</i>	{	putting size	3	16.33			<i>putting</i>	land	7	7.71			
		yard	22	6.59									
		square	9	9.88									
		wide	8	11.75									
		other <i>E</i>	29	5.69									

*Note.—Freak values of r_b exceeding 1 are sometimes obtained if the assumptions underlying the formula for r_b are not wholly justified.

The only *E* answer for which there is any apparent justification is 'size.' It is true that an acre is a measure of size, though this is a much less good answer than 'area.' Only 3 people out of 200 put 'size,' but their mean score was high, nearly as high as $m(R)$. In this case I did not consider the evidence sufficiently strong to justify any action, and the item was retained unaltered. But, if a higher proportion of people had put 'size' or if this had seemed a slightly better answer, some change in the item would have been necessary—see paragraph (c).

TABLE III

Item 14.

"Lighten, 5317246; tile, 2354; hen, ***?"

Correct answer "746": the letter l corresponds to 5, i to 3, g to 1 and so on. This fixes a code which turns 'tile' into 2354 and 'hen' into 746.

Item	<i>R</i>	<i>m(R)</i>	<i>E</i>	<i>m(E)</i>	<i>O</i>	<i>m(O)</i>	<i>U</i>	<i>m(U)</i>	<i>W</i>	<i>m(W)</i>	<i>d</i>	<i>D</i>	<i>r_b</i>
14	47	18.63	54	7.63	91	7.12	8	3.75	153	7.12	11.51	1.658	.971
<i>Analysis of E</i>	{ <i>putting</i> 176		18	11.06			{ <i>putting</i> cock		3	1.66			
	egg		2	5.50			pen		6	3.83			
	<i>other E</i>		25	6.96									

29 different *E* responses were put to this item, mostly odd combinations of digits. The most popular and best *E* answer was '176,' which is obtained by eliminating the digits 2354 from the original digits 5317246. These *E* subjects missed the relevance of the order of the digits, but their answer showed insight into some of the material of the item, though not all of it. 'egg' and 'cock' are amusing answers, presumably suggested by 'hen'; 'pen' simply rhymes with 'hen.' This item was a very good one, as shown by the high r_b of .971.

These examples illustrate the variety of answers that may be given to an inventive-response item and the desirability of careful consideration of each *E* answer in order to decide whether there is any logical justification for it. With a limited-response

test, the number of different E answers is of course limited by the number of 'distracters,' usually three or four, but it is important that each of these should be investigated separately.

(c) *Conditions for scrapping items.* An item is a bad item and must be altered or scrapped if there is an alternative answer which either :

(i) takes account of all the data and is logically justifiable : this holds even if only one experimental subject has put that answer. (To allow alternative answers to an item not only complicates the scoring but is unfair to the high-grade subject who thinks of both answers.)

(ii) while not logically justifiable, has been put by an appreciable proportion of high-grade subjects, so that $d(E)$ is small or negative.

A case which is difficult to decide is when an E answer which is not logically justifiable has been put by a small number of high-grade subjects, so that their mean criterion score is as great or nearly as great as $m(R)$ but the overall values of d and $d(E)$ are satisfactorily high. In my opinion such cases should be treated on their own merits, and it is not always necessary that these items should be altered or deleted. Item 13 in paragraph (b) was an example of an item which it seemed justifiable to retain unaltered.

(d) *Choosing the best items.* The relative value of different items should not be assessed by means of the index d without also considering the difficulty of the items, since d (or D) tends to favour "unbalanced" items, i.e., those relatively remote from 50 per cent. difficulty, whereas tetrachoric r and r_b are (theoretically) independent of p . Items of the same validity, as assessed by r_b , would have values of d or D in the following ratio :

p	..	..	.50	.33 or .67	.25 or .75	.10 or .90	.05 or .95
d or D	...	1.00	1.00	1.02	1.06	1.22	1.36

The effect of item difficulty on d is not very serious since few items for which p is less than .1 or greater than .9 are normally included in a test. Moreover, when an internal criterion (total test score) is being used for item-analysis, if there is any concentration of items of medium difficulty, these items tend to be favoured relative to unbalanced items ; this factor tends to counterbalance the tendency of d to favour unbalanced items, and in practice the net effect is often small.

In any case, even if an index such as r_b is calculated which is (theoretically) independent of p , the difficulty of items should be considered simultaneously with an estimate of their value when choosing the better items from a draft test. The most convenient procedure is to settle first how the items should be distributed according to difficulty in the revised test. It may be decided, for instance, with an 80-item test that items should be distributed evenly according to values of R , ten per decile, from the tenth to the ninetieth percentiles. In this case the items should be grouped by deciles, and within each decile those with the highest values of d retained. Normally some such procedure as this is sufficient : no further calculations need be made. If time permits, however, or if there is little to choose between certain items, it is convenient to plot all items on a graph of d against R and to experiment by drawing a pencil line across the graph, dividing items above the line (to be retained) from those below the line (to be omitted). It is then easy to pick out the better items of appropriate difficulty from those which are less good. The choice of items should not, however, depend solely on d and R . Sometimes an item with high d may have some slight weakness such as a fairly plausible E answer, so that an item with rather lower d but free from any weakness would be preferred. Moreover the content of the items must of course be taken into account as well as their validity indices so that the proper balance of the test is not disturbed in the process of revision. The d method of analysis provides accurate and detailed statistical information about an item, but,

like any other method of item-analysis, it should be applied flexibly, taking account of all available information, subjective as well as objective.

(e) *Two practical techniques.* The process of obtaining full item-analysis data from the test sheets is a lengthy one, and it is important that the various operations should be accurate but as economical of time and energy as possible. Two practical techniques, that of scattering scores on prepared proformas and that of sorting test sheets before making entries, have been developed. They are described by Mr. R. F. Dowse in a duplicated memorandum, copies of which can be obtained on application to the Civil Service Commission Research Unit.

(f) *d and d(F).* Given sufficient extra time for the experimental try-out of a draft test, U will be small; R and d are then reliable indices for all items. If, however, sufficient extra time could not be allowed, U for later items may be large. In such cases, the most reliable index of the easiness of an item is $50(R/N + R/T)$, and the most reliable index of its value is $\frac{1}{2}[d + d(F)]$.

IV. TECHNIQUE FOR EXAMINING DOUBTFUL ITEMS

(a) *Description of method.* When making an item-analysis of an inventive-response test, each different E answer must be separately investigated, as illustrated by the three examples in paragraph III(b). With a choice-response item, each distracter should, ideally, be put by an equal number of subjects, and the values of $m(E)$ should also be equal. In practice, only if the distracters have been selected from the popular wrong answers put to the item when given an initial try-out in inventive-response form, is this ideal likely to be approached even approximately. After the first full-scale try-out of a choice-response test nearly every item can be seen to be capable of improvement; and the first concern of item-analysis should be to guide and facilitate these alterations. Sometimes a distracter must be eliminated because it is too popular or $m(E)$ is too high. More often, one or more distracters are found to be put by so few subjects as to be almost totally ineffective, and should be replaced by more plausible distracters in the second experimental try-out. Such cases are comparatively straightforward. The difficulty arises when m for a particular E is intermediate between $m(R)$ and $m(W)$ and it is not clear whether the item would be improved by allowing that E as correct (if it is logically defensible) or deleting it (if it is not logically defensible) or whether the item would be better unaltered. To help decide these doubtful cases, I have developed the following technique:

In cases of doubt about a particular E , try the effect of transferring these E from W to R . When total test score is being used as criterion, m for these E must be increased by one, as their total score would have been one greater if their answer to this item had been allowed. Then if r_b for the amended item x' is higher than r_b for the original item x , the item would appear to benefit by the alteration, and some action must be taken. If r_b falls, no alteration is needed. It will be appreciated that if p for x is nearer to .5 than p for x' (i.e., if the item was an easy one), only if d does increase need r_b be calculated. For if d does not increase on alteration, neither can r_b . Similarly, if p for x' is nearer to .5 than p for x (i.e., if the item was a hard one), r_b must be calculated only if d does not increase.

(b) *Examples of its use.* In the Army Arithmetic Test SP3B Part 2, question 10 was, "Write .75 as a fraction." Item-analysis results, $N = 524$, were as follows:

TABLE IV

Item	<i>R</i>	<i>m(R)</i>	<i>E</i>	<i>m(E)</i>	<i>O</i>	<i>m(O)</i>	<i>U</i>	<i>W</i>	<i>m(W)</i>	<i>d</i>	<i>D</i>	<i>r_b</i>
10	266	54.51	93	39.22	153	31.29	12	258	33.61	20.90	1.35	.84
	<i>putting</i> $\frac{75}{100}$		27 48.07									
10'	293	54.02	66	35.75				231	31.92	22.10	1.43	.89

27 subjects put 75/100 instead of the correct answer $\frac{3}{4}$. They apparently understood what was meant by a fraction but failed to simplify 75/100 to $\frac{3}{4}$. Their mean score 48.07 was intermediate between $m(R) = 54.51$ and $m(W) = 33.61$, and it was not clear whether 75/100 should be allowed or not. Applying the test, d rises from 20.90 to 22.10 and r_b from .84 to .89, indicating that some action should be taken. The alternatives would be to allow a mark for 75/100 or to rephrase the question, "Write .75 as a fraction in its simplest possible terms" and allow a mark only for $\frac{3}{4}$.

The same test can be applied to items 3 and 14 in the draft GC test Part C referred to in paragraph III(b).

TABLE V

Item 3. "trend, rend, end; scare, care, ***" (Answer "are").

Item 14. "lighten, 5317246; tile, 2354; hen, ***" (746).

Item	<i>R</i>	<i>m(R)</i>	<i>E</i>	<i>m(E)</i>	<i>O</i>	<i>m(O)</i>	<i>U</i>	<i>m(U)</i>	<i>W</i>	<i>m(W)</i>	<i>d</i>	<i>D</i>	<i>r_b</i>
3	82	15.95	87	6.04	30	4.33	1	2.00	118	5.57	10.38	1.500	.933
	<i>putting</i> air		8 9.87										
3'	90	15.50	79	5.66					110	5.26	10.24	1.475	.931
14	47	18.63	54	7.63	91	7.12	8	3.75	153	7.12	11.51	1.658	.971
	<i>putting</i> 176		18 11.06										
14'	65	16.83	36	5.92					135	6.60	10.21	1.470	.897

With item 3, applying the test, r_b is practically unchanged, which suggests that no change is strictly necessary. It would, however, clearly be an improvement to use a sequence of words such as drink-rink-ink with a continuity of rhyme and so avoid any ambiguity.

With item 14, applying the test, r_b falls appreciably, indicating that no change is necessary.

The examples show the usefulness of this technique in helping to make up one's mind about doubtful cases. It should be stressed, however, that the recalculated value of r_b should be used only as a piece of additional evidence to be combined with the other evidence available.

V. FIFTEEN TYPES OF FAULT IN ITEM CONSTRUCTION

With most types of material, it is necessary to allow for a considerable wastage of items in the initial try-outs of a new test. The most likely points of weakness,

which are almost unavoidable, are :

(a) with an inventive-response test, the putting of an ingenious alternative answer by some high-grade subjects.

Example 1. "INSIGNIFICANT TRI" Out of 103 subjects, 25 with $m = 70.36$ put the intended answer TRIVIAL, but three subjects with $m = 77.0$ put TRIFLING. This is just as good an answer as TRIVIAL, and the item was dropped.

(b) with a choice-response test, unevenness of response to different distracters.

Apart from these common points of weakness, I have found from my experience 15 ways in which a test constructor may go astray, if he is not constantly on his guard, and produce items which give disappointing results. These types of item are illustrated by examples from the try-out of draft tests by this Research Unit. An H will be used to denote that the subjects were of Higher School Certificate standard or better.

Type 1. Association of ideas outweighs the logical connexion.

Example 2. "OBSTINATE STU" $N = 103$. $m = 49.70$. 33 subjects with $m = 54.50$ put the correct answer STUBBORN, but 10 subjects with $m = 55.0$ put STUPID, the most familiar word beginning with STU. In a verbal test of this kind it is best to avoid clue letters which strongly suggest some other word. Similarly in a choice-response item it is best to avoid a distracter which is strongly associated with the clue word, e.g. Knife-Fork.

Type 2. Alternative logical connexions.

(a) leading to the same answer.

Example 3. "egg, G ; between, E ; lackadaisical,*". This item gave good results ; but it suffered theoretically at least from the disadvantage that a person might put the correct answer A as being the second letter of the word, rather than by noticing (as intended) that there are 2 G's in "egg," 3 E's in "between" and 4 A's in "lackadaisical."

(b) leading to a different answer. This is a common fault and difficult to avoid with certain types of material.

Example 4. "ruler, L ; pencils, C ; compass,*". $N = 200$ H . $m = 54.54$. 151 subjects with $m = 55.34$ put the intended answer P (the middle letter of the word), but 30 subjects with $m = 55.77$ put A, presumably because L is the third letter in "ruler," C the fourth letter in "pencils," and A the fifth letter in "compass."

Type 3. All-or-none response.

Questions of the "paragraph completion" type tend to suffer from this fault. It is too much a matter of luck whether the subject hits on the right line of thought and scores 9 or 10 out of 10, or works at a wrong line of thought and scores 0. This is particularly true of a test with a strict time limit. It is preferable to use three or four questions of the "sentence-completion" type.

Type 4. Catch distracters.

In an item which asks "Which of the following words means the Opposite of the key-word ?" it is better not to include among the distracters a word which means the same as the key-word.

Sometimes a distracter looks as if it were associated with the key word but in fact is not. This may mislead some of the quicker workers.

Example 5. "SHREWD is to WISDOM as SALUBRIOUS is to ; 1 Salt, 2 Health, 3 Quality, 4 Stupidity ?" $N = 200$ $H. m = 136.21$. 170 subjects with $m = 138.71$ put the correct answer 2, but 9 subjects with $m = 143.22$ put 1. Doing the test at speed, they had carelessly associated 'Salubrious' with 'Salt.' This distracter was intentionally included in order to catch low-grade subjects, but instead it caught comparatively high-grade subjects and so did more harm than good.

Type 5. Colloquial distracter too plausible.

Example 6. "Decide which sentence is better than the others in the group.

- (1) I met him only yesterday for the first time.
- (2) I only met him for the first time yesterday.
- (3) It was only yesterday as I met him for the first time."

(1) is a more grammatical expression than (2) of what is really meant, namely that "yesterday was the day on which I met him for the first time," but (2) is how most people would probably express themselves in conversation. $N = 63$. 26 subjects with $m = 43.35$ put (1), and 36 subjects with $m = 41.25$ put (2).

Type 6. Double meaning of key word.

Example 7. "What word rhymes with CAPE and means the same as FORM ?" $N = 103$. $m = 49.28$. Only 18 subjects with $m = 57.44$ put the correct answer "Shape." As many as 70 subjects with $m = 48.66$ omitted the item altogether. Six subjects with $m = 33.83$ put "Tape." This provided a clue as to what had happened. The O subjects, being Civil Servants, had evidently thought of FORM in the sense of a sheet of paper, and not in the intended sense of the outline of an object.

Type 7. Grammatical errors.

Example 8. "PLEASURE ENJ." $N = 103$. $m = 49.28$. 29 subjects with $m = 55.38$ put the correct answer ENJOYMENT, but 72 subjects with $m = 46.88$ put ENJOY, not bothering to put the correct grammatical form. Similarly in a choice-response item it is undesirable for two distracters to have the same general meaning and differ only in grammatical form.

Type 8. Inadequate clues.

In an item of the ordinary series type sufficient terms must be given to minimise the chance of finding ingenious alternative methods of continuing the series.

Example 9. "3, 4, 6, 9, 13, . . . ?" $N = 144$ $H. m = 140.45$. 121 subjects with $m = 141.92$ put the intended answer '18, 24,' continuing the series with successive terms increasing by 1, 2, 3, 4, 5 . . . Three subjects with $m = 136.33$, however, put '19, 28': they had hit upon the ingenious idea of adding to the previous term the term two places to the left of that— $3 + 6 = 9$. $4 + 9 = 13$. $6 + 13 = 19$.

Alternative layouts which have the advantages of being more likely to avoid ambiguity and of providing the subject with a check on the accuracy of his working are :

- (a) "3, 4, 6, 9, 13, . . . , . . . 31. What are the two missing terms ?"
- (b) "3, 4, 6, 9, 14, 18, 24, 31. Which is the incorrect term ?"
- (c) "3, 4, 6, 9, 11, 13, 18, 24. Which is the superfluous term ?"

Type 9. Level of population inappropriate for the item.

It is interesting to compare item-analysis results for the following "Implications" item as tried out on two different levels of population.

Example 10. "Which two of the numbered words are necessarily implied by TO FORGIVE? : 1 Contrition, 2 Friendliness, 3 Injury, 4 Forget, 5 Past?" Correct answer, 3 5. One can forgive a past injury without e.g., requiring contrition, exhibiting friendliness or promising to forget the injury.

Tried out first on 89 *H* subjects with $m = 76.06$, the item worked well. R was 27, $m(R)$ 88.52; E was 62 and $m(E)$ 70.63. $d = 17.89$.

When tried out in a shorter test, however, on 200 ordinary subjects, the item gave poor results. R was 9, $m(R)$ 17.77; E was 141 and $m(E)$ 11.16. $d = 8.31$. $r_b = .52$. The subjects who put 3 4, 1 3, 2 3 and 1 5 obtained means higher than $m(R)$. This item was clearly inappropriate at this level, though other items of the same type worked well.

Type 10. Obscurity of key word (in a test other than a vocabulary test).

Example 11. "Which two of the following six words mean the same, or nearly the same? 1 precept, 2 obedience, 3 command, 4 request, 5 search, 6 question." $N = 264$ *H*. $m = 50.98$. 62 subjects with $m = 52.50$ put the correct answer 1 3, but 50 subjects with $m = 51.90$ put 4 6, and 63 subjects with $m = 50.51$ put 3 4. The majority of these subjects probably did not know what "precept" meant.

Type 11. Position of most plausible distracter too early.

Example 12. "Which two of the following six words are opposite, or nearly opposite in meaning? 1 calm, 2 energetic, 3 rash, 4 brave, 5 prudent, 6 stupid." $N = 264$ *H*. $m = 50.98$. 183 subjects with $m = 52.76$ put the correct answer 3 5, but 42 subjects with $m = 49.05$ put 1 3: allowing this answer as correct, d increases from 6.1 to 8.1. If 1 and 5 had been interchanged, to ensure that all subjects read the word prudent before the word calm, the item might have given satisfactory results.

The most plausible distracter cannot of course always come after the correct answer, as it is necessary for both to be distributed more or less evenly over all possible positions. It is therefore a good rule to put the most plausible distracter after the correct answer in items where it is distinctly plausible and before the correct answer in items where it is less plausible.

Type 12. Shades of meaning 'too fine.

Example 13. "Which two of the following six words mean the same, or nearly the same? 1 clever, 2 content, 3 proficient, 4 thoughtful, 5 expert, 6 veteran." $N = 200$ *H*. $m = 54.54$. 104 subjects with $m = 56.61$ put the intended answer 3 5, but 31 subjects with $m = 56.35$ put 1 5, and 53 subjects with $m = 50.36$ put 1 3. "Clever" is rather too similar in meaning to "proficient" and "expert."

Type 13. Unclear presentation of problem.

Example 14. "introduce, 123456789; tune, ****." $N = 264$ *H*. $m = 50.98$. 139 subjects with $m = 54.64$ put the correct answer 3729, but 61 subjects with $m = 51.16$ put 1234: counting this answer as correct, d increases from 7.85 to 11.39. The item would be improved by rearranging the digits 1 to 9 in random order and providing an additional clue—compare the successful item 14 described in paragraph III(b).

Type 14. Unusual relationships.

Example 15. "173 119 146 8*." $N = 264$ *H*. $m = 50.98$. Correct answer 3 ($8 + 3 = 11 = 1 + 4 + 6$ etc.) R was 123 and $m(R)$ 52.24; but E was 74 and $m(E)$ 51.54. Most of the E subjects probably took the numbers to represent a series, which was a natural train of thought.

Type 15. Wording not clear.

Example 16. "Mary and Jane are blondes. Ethel and Bertha are brunettes. Only Jane and Ethel are tall. Who is the short blonde: 1 Mary, 2 Jane, 3 Ethel, 4 One cannot say?" $N = 200$. $m = 166.16$. 164 subjects with $m = 167.80$ chose the intended answer 1. 26 subjects with $m = 159.50$ marked '4 One cannot say.' They doubtless argued that 'short' and 'tall' are not mutually exclusive terms; a person can be neither short nor tall but of medium height. Hence there might not be any short blonde. This example illustrates the importance, particularly in a Reasoning problem, of making the wording as watertight as possible.

The reasons why these 16 items did not work well become clear enough once the item-analysis results have been seen. It is doubtful, however, whether the average test constructor would have foreseen more than half of the snags beforehand. Certain types of fault can be avoided as one becomes more experienced in test construction, but others deriving from e.g., ambiguity of language are hard to avoid. Thorough item-analysis is an indispensable pre-requisite to producing a reliable and valid test.

VI. SUMMARY

A method of item-analysis has been described which makes full use of information derived from experimental trials, with a view to identifying faults in items and, if possible, correcting them. The best items should be chosen taking account of item difficulty as well as item validity. Fifteen types of weakness in an item are given, with examples.

My thanks are due to the Civil Service Commissioners and to the War Office for permission to quote examples from experimental investigations. The opinions expressed in this article, however, are entirely my own and are not necessarily shared by the Civil Service Commissioners or by the War Office.

THE FACTORIAL STUDY OF TEMPERAMENTAL TRAITS

By CYRIL BURT

Department of Psychology, University College, London

I. *Nature of Problem.* II. *Normal Cases.* III. *Psychoneurotic Cases.* IV. *Summary and Conclusions*

I. NATURE OF PROBLEM

The Basic Hypothesis. The doctrine of temperamental types, in both its ancient and its modern form, rests on the hypothesis that "temperament is largely a matter of bodily constitution," and that, as a result, "emotional differences are closely associated with bodily differences, the latter providing both important causes, and suggestive symptoms, of individual variations in character or disposition" (McDougall, 1, pp. 99f. ; cf. 6, pp. 7, 19). An obvious method of verifying this hypothesis is to calculate coefficients of correlation between the emotional types and the physical types. To do this, however, we must first demonstrate the existence of such types, and provide some operational method for determining to which type each individual belongs or how closely he approximates to it.

Many modern investigators—De Giovanni and his followers in Italy, Sigaud and Macauliffe in France, and more recently Bauer and Kretschmer in Germany—without offering any clear definition of what they understand by a type or any very convincing proofs for their conclusions, have alleged that, among both the normal and the abnormal, such types are readily discernible 'to the eye of the experienced observer,' and that a classification according to physical characteristics may consequently prove an aid to diagnosis. More recently the French and Italian schools have attempted to measure the types, and to study their concomitant variations statistically ; and it has been contended that the striking similarity between the descriptions independently given by so many writers in so many different countries affords at least a strong presumption in favour of the hypothesis in its modern form (cf. Piéron, 12, and discussion ; also tables *ibid.*, pp. 97–99). On the other hand, the claims so frequently made have not passed unchallenged. "It is impossible," wrote McDougall, "to distinguish any clearly defined classes of temperaments, as many modern authors, as well as ancient, have attempted to do : some of the best modern psychologists have been led into absurdities by attempting the task. In one respect only can we advance upon the ancients—we can realize the great complexity of the problem, and frankly admit our ignorance." And this on the whole would seem to be the opinion of most writers at the present day.¹

Methods of Verification. The problem is essentially one of classification ; and therefore lends itself to investigation by means of factorial techniques. In the

¹ Cf. (1), 1908, p. 100. As we shall see, however, in later editions, McDougall appreciably modified his opinion. For a general review of the subject, see Schrieder, E. (1937), *Les Types Humaines* (Paris : Hermann et Cie). Still more recent discussions will be found in Wells, F., *Am. J. Psych.*, LI, 1938, pp. 136–145 ; cf. also Burchard, M. C., *Comp. Psych. Mon.*, XIII, 1936 : both deal chiefly with the mentally abnormal. Gruhle, H. W., *Arch. f. Psychiat.*, LXXVII, 1926, pp. 1–31, deals with *der Körperbau der Normalen*. J. I. Cohen's thesis ('Physical and Psychotic Types,' Univ. London Lib.) gives a survey of the literature down to 1940.

previous number of this journal (p. 98) I pointed out that two simple statistical procedures are available for verifying the underlying assumptions—namely, ordinary factorial analysis and canonical analysis; and I argued that, though both should give essentially the same results, the most satisfactory procedure for testing the crucial issue would be to make separate factorial studies of physical measurements and emotional assessments respectively, and then, if the results supported the hypotheses of physical and of emotional types, to investigate the correlation between the two. From the factorial standpoint a 'type' may conveniently be defined as a hypothetical pattern or profile of factor-measurements towards which the individuals in a given group or class approximate in varying but measurable degrees. The existence and importance of each type can thus be made the subject of precise numerical determination.

The existence of physical types¹ has now been adequately demonstrated (cf. this *Journal*, I, p. 113 and refs.). On the other hand, in spite of numerous plausible descriptions, the existence of emotional types still remains an open issue. "The very attempt at classifying emotions, as we attempt to classify colour sensations or the sensations of touch, is," we have been told, "out of the question, at least in the present state of knowledge."² In this article, therefore, I shall be concerned to consider evidence for the existence and nature of emotional types; and in a later paper I shall discuss the relations between the two sets of type-factors.

Subjects. In earlier enquiries (5, 7, 8, 10), based on comparatively small groups, I attempted a provisional analysis for physical and emotional characteristics both among normal adults and among normal children. However, among adults the frequency of leptomorphic physiques and of inhibited emotional attitudes tends to diminish with age. Consequently, the alleged correlations, reported by Kretschmer and others for adults, might conceivably be due, wholly or in part, to the age-differences alone. Here, therefore, I shall confine myself to children below the age of puberty. The material to be analysed consists of data collected for (a) children referred for psychological examination from the London County Council's schools, and (b) ordinary children, studied as control groups or during the course of special surveys. The majority of the children were between 9 and 13 years of age.

Of the cases sent to me during the course of my official work, more than half were referred wholly or partly on account of what may be called 'orectic' disorders: that is, they were what would ordinarily be termed 'delinquent' or 'neurotic' cases. The majority, just over 500 in all, were subjected to a full and systematic investigation in accordance with a general scheme that has been fully explained elsewhere (10 and refs.). For purposes of our present problem we shall require to study only those cases which were suffering primarily from some type of emotional disturbance. I shall therefore select those that had either been tentatively diagnosed as cases of 'psychoneurosis,' 'personality defect,' or 'nervous trouble' (in the sense of 'functional nervous disorder'), or were found to exhibit symptoms similar to those which might lead to some such diagnosis. This yields a total of 328 cases.

The normal group comprises 483 boys and girls attending ordinary elementary schools. The majority were studied as control groups during investigations of backward, delinquent,

¹ For reasons I have given elsewhere (6, p. 221; 7, p. 186) a factorial study of physiological characteristics (such as individual differences in metabolism, autonomic stability and endocrine balance) would in my view be even more instructive than the study of differences in body-build. Unfortunately our own experience has been that most of the tests of such functions, available for application on a comparatively large scale, possess only a low reliability, and yield very low correlations. The same conclusion has been reached by other investigators in this field (cf. Darling, R. P., *J. Abn. Soc. Psych.*, XXXV, 1940, pp. 246-260; Wenger, M. A., *Psychosom. Med.*, III, 1941, pp. 427-434; IV, 1942, pp. 84-95). Using rather more elaborate techniques, however, Mr. J. Lindsay has recently obtained somewhat more reliable measurements in a research still in progress, which I hope will shortly be published.

² Ruckmick, C., *Psychology of Feeling and Emotion* (1936), p. 120, quoting Külpe.

and neurotic children. So far as relevant characteristics are concerned, they were examined, measured, and assessed in the same way (though not so intensively or in such detail) as the cases referred by teachers, school medical officers, magistrates, or clinics.

Traits. The actual symptoms observed in what may be called the 'psycho-neurotic' group have been described and tabulated in an earlier contribution¹ (9, p. 15). In that paper I attempted a factorial study of *all* the symptoms observed or reported. Here, however, we are concerned solely with a classification according to emotional characteristics. In this article, therefore, I shall limit the analysis to assessments for the 'primary emotions.'

The term 'primary emotions' is used in the sense defined by McDougall (1, pp. 45f.).² According to McDougall, "each of the principal instincts, which man shares with the higher mammals, conditions some kind of emotional excitement, whose quality is specific or peculiar to it." These emotional propensities may therefore be regarded as "innate, elementary, conative tendencies," inherited by all human beings. They "constitute the chief mental forces of which a genuinely *dynamic psychology* has to take account." Since these tendencies are inborn, differences in their strength or intensity may also be regarded as inborn; and it is this that, in my view, makes them of fundamental importance in individual psychology.

With the normal group the assessments for these primary emotional tendencies were based in the first instance on gradings submitted by teachers. The investigations summarized here, however, were undertaken at intervals during a period of over twenty years; and the various teachers and research-workers who assisted at different stages of the work naturally had somewhat varying standards. So far as possible their standards and judgments were checked by a personal and more intensive study of selected cases and selected groups. The children in the psychoneurotic group were all interviewed personally, and assessed by the methods I have described in previous publications.³ In every case the final gradings were expressed as standard deviations from what was assumed to be the average or normal level for a boy or girl of the mental age in question.

Previous Investigations. One purpose of the present enquiry is to confirm and to extend the conclusions that seemed to emerge from earlier studies, undertaken with smaller numbers and rougher methods over thirty years ago. In a tentative enquiry, which was commenced as early as 1911 during my work with school children in Liverpool, an endeavour was made to apply factorial methods to emotional characteristics on lines similar to those already developed for studying cognitive characteristics. My general aim was, first to determine what I called the 'highest

¹ This was based mainly on the earlier L.C.C. *Report on Emotional Disorders in Children of School Age* (1930), from which much of the following data is taken; and I have to repeat my acknowledgments to the Council for permitting me to make use of material gathered while I was in their service. Since the date of that report, however, a number of further cases have been added to my list; and the correlations and factor analyses have been worked out afresh. I am especially indebted to Miss G. Bruce for help with the earlier enquiry and to Dr. C. Banks for assistance with calculations for the following tables.

² In a later work (2, pp. 97f.) McDougall slightly enlarged his list. Here, as in my 1915 investigation, I have limited myself to the short original list (the 'propensities' numbered 2-11 and 14 in his later list). However, a 'comfort seeking propensity' (numbered 15 in the later list) was early added to McDougall's list; and at his suggestion I included it in my studies after 1916. This makes 12 primary traits in all. Other 'propensities,' found in McDougall's fuller lists—acquisitiveness, constructiveness, talkativeness, and the like—were included in the assessments; but they are omitted here, because they have no obvious 'emotional' correlates. One or two names have also been altered to conform with the later nomenclature ('Elation' has been called 'Assertiveness,' and 'Wonder,' 'Curiosity').

³ See 8, 10, and refs. *ibid.*; and for a more intensive study of the reliability and validity of the various procedures, cf. 'The Assessment of Personality,' *Brit. J. Educ. Psych.*, XV, 1945, pp. 107-121.

common factor' for all the traits taken jointly (that is, the component giving the closest fit to all the trait-measurements and correlations, or, as we should now put it, accounting for a maximum of the variance), and then to test the significance of the residuals left after that factor had been subtracted, in order to ascertain whether any further factors were required. The mathematical procedure adopted was based essentially on the method suggested by Karl Pearson in 1901 for the analysis of physical characteristics (3). It differed from Pearson's in two main respects, namely (to use terminology which has since become current), (i) in seeking to determine the *minimum* number of orthogonal factors needed, rather than the *maximum* number, and (ii) in substituting *simple* sums for the *weighted* sums that Pearson's formulæ strictly required.¹ In order to cope with problems of *more* than one common factor it was assumed that the factors in a 'dynamic psychology' might be regarded as combining like the hypothetical components (displacements, velocities, forces, etc.) of ordinary dynamics.

To determine the nature of the factors, it was insisted, both internal and external criteria should be used. The internal criterion consisted of a study of the traits having the highest factor-saturations and the persons having the highest factor-measurements. The external criterion consisted of independent assessments obtained from teachers (or other acquaintances) who graded the persons for the particular qualities with which the factors were tentatively identified. This was supplemented, wherever possible, by introspective reports from the testees.²

The results thus reached were described in a paper submitted to the Psychological Subsection of the British Association in 1915 (5). The analyses, based on data from 172 normal children and 157 normal adults, led to the hypothesis of a common factor underlying all the primary emotions, which was termed 'general emotionality' (by analogy with 'general intelligence'). When this factor had been partialled out, the residual correlations appeared to indicate two bipolar factors, classifying the emotions (a) according to their 'sthenic' or 'asthenic' character and (b) according to their 'algedonic' quality. In a later paper it was noted that these results appeared in some degree reminiscent of the traditional classification of temperaments, and that a study of bodily measurements furnished indications of small but positive correlations between the emotional factors and the physical characteristics supposed to be distinctive of the temperamental types—particularly the leptomorphic (or micro-

¹ It still seems widely supposed that the ordinary 'principal axes solution' remains the 'ideal procedure,' even when reduced communalities are substituted for unities in the diagonal (cf. Kellogg, C. E., *J. Educ. Psych.*, XXVII, p. 581 and Thurstone, L., *Multiple Factor Analysis*, pp. 295-6). But, as I have previously pointed out, if we decide to seek, not a maximum number of factors (n factors for n tests), but a minimum, then, in theory at any rate, the determinantal formula deduced by Pearson must be modified. Instead of his equation $|R - \lambda I| = 0$, we must substitute $|R - \lambda R_c| = 0$, where R denotes the observed correlations with unities in the diagonal and R_c the reduced correlation matrix. Let us write $R = R_c + D$: then, if the reduction affects the variances only, D will be a diagonal matrix containing the specific factor variances. The modified equation can readily be deduced by inserting the supplementary assumptions into Pearson's proof. Alternatively, it can be reached by regarding the observed and the hypothetical traits as two sets of variables, and then using the ordinary formula for the canonical correlation: (see this *Journal*, p. 103, footnote 1 and ref.; also *ibid.*, p. 105, footnote 2).

In practice the change would seldom make much difference to the actual size of the saturations. But the simplified procedure involves two consequences, which are commonly forgotten, and will be noted later on (pp. 185 and 189, footnotes 4 and 2).

² Of late there has been a tendency for factorists to overlook the need for criteria of the second kind. Their importance was most clearly brought out in the discussions following my paper demonstrating the general factor underlying cognitive abilities, and notably in the printed comments by Dr. C. S. Myers (*Brit. Ass. Ann. Rep.*, 1910, p. 807). The difficulty, of course, is that (as happened in the research here described) the nature of the supplementary factors may not become clear until the analysis has been completed. This can best be overcome by making a 'pilot investigation' first of all.

TABLE I. NORMAL CHILDREN: CORRELATIONS BETWEEN TRAITS

Trait	1	2	3	4	5	6	7	8	9	10	11	12
	Joy	Sex	Sociability	Assertiveness	Anger	Curiosity	Disgust	Sorrow	Fear	Submissiveness	Tenderness	Comfort
1. Joy ..	(.867)	.339	.841	.577	.473	.461	.289	.631	.452	.192	.535	.336
2. Sex ..	.339	(.171)	.401	.271	.237	.229	.103	.257	.103	.052	.211	.089
3. Sociability ..	.841	.401	(.998)	.753	.719	.577	.313	.617	.282	.030	.491	.274
4. Assertiveness..	.577	.271	.753	(.697)	.670	.432	.252	.346	-.005	-.094	.266	.157
5. Anger ..	.473	.237	.719	.670	(.995)	.532	.326	.585	.077	-.167	.203	-.045
6. Curiosity ..	.461	.229	.577	.432	.532	(.390)	.182	.410	.142	-.045	.255	.073
7. Disgust ..	.289	.103	.313	.252	.326	.182	(.188)	.351	.236	.086	.182	.087
8. Sorrow ..	.631	.257	.617	.346	.585	.410	.351	(.862)	.683	.221	.458	.196
9. Fear ..	.452	.103	.282	-.005	.077	.142	.236	.683	(.854)	.438	.428	.230
10. Submissiveness	.192	.052	.030	-.094	-.167	-.045	.086	.221	.438	(.311)	.207	.169
11. Tenderness ..	.535	.211	.491	.266	.203	.255	.182	.458	.428	.207	(.380)	.244
12. Comfort ..	.336	.089	.274	.157	-.045	.073	.087	.196	.230	.169	.244	(.229)

splanchnic) and euryomorphic (or macrosplanchnic) types of De Giovanni, Viola, and their Italian fellow-workers (cf. 6 and 7).

The existence of a general factor on the conative side, analogous to 'general intelligence' on the cognitive side, has been confirmed by a number of subsequent writers, such as Webb, Oates, and Cattell (14, 15, and refs.). But the factor which has been most frequently verified is the larger of the two bipolar factors. Under various names—'sthenic and asthenic,' 'unrepressed and inhibitive,' 'extravertive and introvertive,' 'surgent and desurgent,' 'cyclothymic and schizothymic'—this factor has appeared and reappeared in numerous investigations and speculative classifications. In this country the best-known study of the relations between emotional and physical characteristics is still that described in Kretschmer's *Physique and Character* (1925); and the distinctions there discussed would seem to turn largely on the presence of this second factor in his pathological groups. The third factor (the algedonic or euphoric-versus-dysphoric factor) has found a much smaller degree of support. The contrast between the 'sanguine' and the 'melancholy' temperament—*l'allegro* and *il penseroso*—is familiar enough in literature; but the antithesis seems more frequent in popular writings than in psychological.

The conclusions drawn in these earlier papers have not escaped criticism. Prof. J. A. Green "doubted whether either intelligence or emotional behaviour could be treated as a multidimensional entity capable of 'resolution' in certain directions."¹ Others, like C. S. Myers, were "prepared to admit the applicability of mathematical techniques to measurements by cognitive tests, but thought them inappropriate in the emotional or conative sphere." Spearman, here as elsewhere, disputed the assumption of group or other supplementary factors. American writers have doubted the existence of the general factor. Nunn "believed that the principle of the combination and resolution of hormic forces might be suggestively applied to certain actions of an individual, but questioned its legitimacy in dealing with crude assessments for a random batch. Might not different batches " (he asked) "yield very different results?"

II. NORMAL CASES

Correlations. My first object here, therefore, will be to meet these criticisms, so far as I can, by applying the same procedure to "different batches," and, indeed, much larger batches, and comparing the results with those previously obtained. We may begin with the 483 normal or ordinary school children described above. The correlations² between the various traits are shown in Table I. The coefficients are on the whole a little smaller than those reported in my earlier investigation: the reduction is doubtless due to the variations in standard referred to above.

In its general character the table closely resembles the first of the tables printed in my 1915 paper (5, p. 695, "Table I: Observed Coefficients for Children aged 9-12"). With few exceptions all the correlations are positive; and the coefficients show a rough approximation to a hierarchical arrangement: but, as in the earlier table, the hierarchical arrangement is broken by an irregular ridge of comparatively high correlations which appear on either side of the leading diagonal, and suggest a kind of cyclic overlapping between the successive traits.

Factor Analysis. With these numbers, and with correlations of this moderate size, not more than four factors are likely to be significant; and, as it turns out, a tolerably good fit can be obtained with only three. Four factors have been extracted by the same formula as before; and the saturations are given in the first four columns of Table II.

¹ Memorandum to Brit. Ass. Committee on 'Mental and Physical Factors in Education' (1919). Later, however, he leaned towards a more sympathetic view, but "still felt the criticism had some force" (cf. his memorandum in the Board of Education *Report on Psychological Tests of Educable Capacity*, 1924, p. 230).

² The correlations were calculated separately for the different groups, and then averaged by weighting each table in proportion to the number in each group. The original correlations were given in my L.C.C. *Annual Reports* as each enquiry was completed. The figures in brackets are the communalities for the first five factors: (to make sure that the n th factor is devoid of significance, it is generally wise to include the squares of the saturations for the $(n+1)$ th).

TABLE II. NORMAL CASES: FACTOR-SATURATIONS AND CORRELATIONS WITH CRITERIA

Trait	Internal Criterion				External Criterion			Normalized Saturations	
	I	II	III	IV	I	II	III	II	III
1. Joy	.876	.032	.250	.068	.590	.390	.610	.128	.992
2. Sex	.360	.117	.104	.106	.223	.177	.098	.747	.664
3. Sociability ..	.921	.343	.148	.047	.632	.422	.521	.918	.397
4. Assertiveness..	.632	.491	.090	— .221	.496	.622	.218	.983	.181
5. Anger.. ..	.673	.525	— .495	— .142	.611	.584	— .381	.727	— .687
6. Curiosity ..	.532	.278	— .097	.142	.311	.426	.101	.944	— .331
7. Disgust ..	.379	— .029	— .123	— .168	.215	— .106	— .097	— .226	— .974
8. Sorrow ..	.821	— .244	— .338	.055	.533	— .247	— .543	.585	— .811
9. Fear	.573	— .694	— .142	.113	.417	— .634	— .468	.980	— .201
10. Submissiveness ..	.205	— .475	.129	— .062	— .162	— .597	— .112	.965	.262
11. Tenderness ..	.564	— .164	.152	.107	.246	— .263	.233	.733	.680
12. Comfort ..	.301	— .180	.322	— .045	— .193	— .182	.354	.486	.873
Sum of Squares	4.479	1.553	.652	.169	2.141	2.204	1.742	—	—
Contribution to Variance..	37.3%	12.9%	5.4%	1.4%	—	—	—	—	—
Column	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(ix)	(x)

The general nature of the first three factors seems clear. The first has positive saturations throughout ; and may therefore be identified with the factor for 'general emotionality.' It contributes over 37 per cent. to the total variance. As in the previous research (*loc. cit.*, Table I), I have secured from the teachers a set of independent gradings for 'general emotionality' ; and, as before, the close correspondence between (a) the correlations of the several traits with this external criterion (col. vi) and (b) the correlations of the same traits with the internal criterion, supplied by the general factor itself (col. i), affords strong evidence for the identification. Although the correlations with the external criterion are (as usual) somewhat lower than the correlations with the hypothetical factor, the proportions are much the same. The amount of agreement may be assessed by what I have termed the 'proportionality criterion' : this is obtained by calculating the 'unadjusted correlation' between the two columns

of coefficients viz., $\frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = 0.931$. The most discrepant cases arise in the latter half of the table : the saturations for these traits are disproportionately larger than their correlations with the teachers' assessments. No doubt teachers are apt to underestimate the emotionality of the submissive or unhappy child, who does not show his feelings in so conspicuous a fashion as the more aggressive.¹

The second factor is bipolar, and divides the emotions into a 'sthenic' (or demonstrative) group and an 'asthenic' (or inhibitive) group. It contributes 13 per cent. to the total variance. This factor appears to afford some confirmation of the doctrine of introverted and extraverted types. However, the distribution of the factor-measurements, which, as I have shown elsewhere,² is approximately normal, makes it clear that the so-called 'types' are merely the extreme tail-ends of a normal distribution. To check the interpretation of the factor the teachers were asked to furnish independent ratings for the degree to which the children exhibited or repressed their apparent feelings. In this case the agreement (cf. cols. ii and vii) is even closer : the proportionality criterion yields a figure of 0.967.

The third factor contributes 5.4 per cent. to the total variance. At the time of the investigation it was identified with 'cheerfulness' or 'optimism' (versus the opposite). Taking teachers' ratings for this quality (col. viii), the proportionality criterion yields a figure of only 0.783. Hence this proposed interpretation of the factor seems a little too crude : probably it would be more accurate to say that that factor divides the emotions into those that are accompanied by pleasurable and by unpleasurable feeling-tone respectively.³

The fourth factor contributes only 1.4 per cent. to the total variance. It is below the borderline for statistical significance ; and, as we shall find in a moment, is probably attributable to the inclusion of a few abnormal or pathological cases.⁴

¹ If we require *internal* evidence that the general factor is essentially an emotional or conative factor, and not some still more comprehensive factor common to *all* mental processes (as Spearman at one time contended), our selection of traits should include cognitive abilities as well as emotional propensities. When that is done, the factor proves to have approximately zero saturations for the former, and the 'general factor' becomes converted into a 'group factor.'

² *Factors of the Mind*, p. 432, Table VI.

³ For a fuller discussion of the nature of these factors, see 8, pp. 285-9.

⁴ As our analysis is not based on the full modified formula (p. 181, footnote 1), and as the initial correlations themselves are averages, the ordinary chi-squared test is not strictly applicable. It shows a very large rise in *P* after the third factor has been extracted. This may be due to the calculations just mentioned ; but to my mind it rather suggests that the gradings for the several traits were not made quite as independently as was desirable. However, these doubts cannot affect our judgment as to the significance of the factors.

It will be seen that, if we assume that the twelve traits provide a sufficiently complete or a sufficiently representative sample of the particular field of characteristics with which we are concerned, namely, that of the 'primary emotions,' the factors extracted from them by the summational procedure (including the bipolar factors with their negative saturations) can be given a concrete psychological interpretation just as they stand. 'Rotation' is unnecessary. This would seem to be one of the main points of difference between myself and most American investigators who have since used much the same mathematical procedure.¹

The Classification of Emotions. Together these three orthogonal factors, as I pointed out in the original discussion, are strongly suggestive of Wundt's well-known tri-dimensional theory of the feelings. According to Wundt "the entire system of the feelings can be regarded as a three-dimensional manifold, in which each of the three dimensions includes two opposite directions."² Any complex state of feeling may thus be resolved into three abstract *Faktoren oder Komponenten* (p. 318); and these *Gefühlskomponenten* are regarded as combining to yield composite *Gefühlsresultante*, after the analogy of the fusion of sensations in colour-mixture.

As applied to the emotions, this hypothesis leads to a threefold cross-classification,³ not unlike that set out above. Thus (i) Wundt begins by arranging all the emotions in a serial and bipolar scale according as they are more or less pleasant or unpleasant: this dimension (*Lust* versus *Unlust*) corresponds with our third factor. (ii) He then introduces a cross-division according to the degree of excitement (*Erregung*) or inhibition (*Hemmung*, *Deprimierung*, *Beruhigung*): this is analogous to our second factor. (iii) His third 'dimension'—*Spannung* versus *Lösung* (usually translated 'tension' and 'relaxation')—is related to the relative intensity of the feelings and to their stability or instability as manifested by their temporal course: the more violent and sudden emotions are assigned to one end of this third scale, and the weaker and more gradual responses to the opposite end: here there is some resemblance to the order of emotions as arranged by our first factor, although the principle of classification is no longer quite the same.

The several components may combine in varying amounts. Thus anger is a fusion of high excitement with marked unpleasure; fear, of high tension with marked unpleasure; hope, of high tension with marked pleasure. Wundt insists that the combination is no mere *additive Verknüpfung der Elemente*, no mere summation of *Partialgefühle* to produce *Totalgefühle*: the analogy is rather, he says, with the resolution of a motion in space according to its three co-ordinates, or better still with the three-dimensional system formed by visual sensations, as illustrated in the familiar *Farbenkugel* or colour-pyramid.⁴

The Classification of Persons. As applied to the problem of individual differences, the three factors which we have obtained lead to a graded classification of persons according to what Shand and McDougall proposed to call their 'tempers.' According to Shand,

¹ Cf. 'Symposium on Factorial Analysis,' *Brit. J. Psych.*, XXX, p. 91 and refs. If we wish, we may, of course, seek further light by rotation or some equivalent method of substituting group factors for bipolar; and, in dealing with cognitive abilities as distinct from temperamental traits, it is (in my view) nearly always desirable to do so. But that is very different from declaring (as is usually done in descriptions of the Thurstone method) that "centroid axes usually lack psychological meaning" (Guilford, *Psychometric Methods*, p. 488, heading to first paragraph on 'Rotation').

² *Grundzüge der Physiologischen Psychologie*, 1910, II, p. 298, and Fig. 230.

³ *Ibid.*, III, p. 211.

⁴ *Loc. cit.*, p. 353. By 'addition' or 'summation' Wundt here means numerical or algebraic addition. As his reference to the resolution of motion in physics implies, his 'resultants' could in theory be obtained by vectorial 'addition,' though Wundt nowhere explicitly suggests this. (The distinction can be illustrated by the addition of two notes, say C and E, which produces not a third note but an audible mixture or 'chord,' as contrasted with the addition of two colours, say Red and Green, which produces not a visible mixture of Red-plus-Green, but a third intermediate colour, Yellow.) Wundt's scheme for the classification of the emotions is, of course, intended to cover a far wider range of psychological phenomena than is mentioned in the text: I have merely picked out the points that are relevant here. An attempt at verifying the further details of this scheme by a more elaborate statistical enquiry was made in a laboratory research undertaken at my suggestion by Miss A. W. Moore; but the groups were too small for definite conclusions to be drawn.

"a man's temperament is the sum of the innate tempers of his different emotions";¹ and McDougall, in a later edition of *Social Psychology* (1919, pp. 380f.), discusses how far the 'innate tempers' as exhibited by different individuals may be classified in terms of three 'independent variables.' "The principal factors of temper," he concludes, "seem to be of three kinds. First, the conative tendencies may all be strong or all be weak. Secondly, they may be either extremely persistent or but little persistent. Thirdly, a greater factor of temperament—also independently variable—is the native susceptibility to pleasure or pain. The conational or emotional tendencies may thus be low or high in each scale independently of its position in the other. . . . If the conative endowment of individuals varies in these three ways, we can explain all the varieties of temper as conjunctions of different degrees of these three attributes. There will consequently be eight well-marked types, corresponding to the eight possible combinations: " that is to say, we have $2 \times 2 \times 2 = 8$ extreme temperamental types, resulting from the crossing of three bipolar factors, and between these eight extremes all individuals may be supposed to vary. Here, it will be seen, McDougall adopts the essential concepts of factor analysis as applied to conational psychology, and particularly to the problem of temperamental types; and appears in the main to accept the conclusions drawn from my earlier investigation.²

The Graphical Representation of Factors. The familiar 2×2 classification of temperamental types would appear to rest essentially upon the second and third factors in Table II above. In order to exhibit more clearly the relations arising from this pair, I attempted in my former paper to represent the numerical results by a simple two-dimensional diagram, based on conventions commonly employed in mechanics. The actual method of construction was suggested by the circular diagrams sometimes used by Karl Pearson to exhibit the correlations between two or more variables.³ In these the degree of concomitant variation between any pair of traits is represented

¹ *Foundations of Character* (1914), p. 129.

² I should add that my own investigation owed much to his suggestions and encouragement, especially during the years when I was working in his laboratory (1904-9). Although he himself undertook no statistical researches, he had from the very outset appreciated the importance of adopting the Galton-Pearson idea of "reducing anthropometric characteristics to terms of uncorrelated or independent variables," both in the cognitive and in the conative sphere.

In the passage quoted above, since McDougall is thinking not of factor-saturations for traits, but of factor-measurements for persons, there is no inconsistency in treating the 'general factor' (strong or weak general conative energy) as bipolar: all factor-measurements, being expressed as deviations above or below the mean, are bipolar. He does not, however, seem to have observed that his scheme of $2^3 = 8$ temperaments is almost exactly identical with that of Höfding, *Psychologie in Umrissen* (2 Aufl., pp. 475-488), a textbook which he himself often recommended.

In regard to the points discussed in the text, there was at first some slight divergence between us over the nature of the second factor. I had stated that it seemed suggestive of "the dichotomy between the 'restrained' and the 'unrestrained' types, the 'sensitive' and the 'excitable,' or (in terms adopted by Wundt) the 'subjective' and the 'objective'" (cf. 6, p. 22). McDougall wrote in reply: "I fully agree with the attempt to reduce such variations to two or three independent variables or 'factors.' But, if the Wundtian dimensions are to be invoked, I should interpret your second factor as corresponding with the dimension of 'persistence' or 'stability'—terms that to my mind, however, are more suggestive of Müller's 'Perseveration' than of Wundt's 'Spannung.' Moreover, I should regard the 'tempers' as describing not so much the primary emotions themselves, but rather the way each or all of them may work towards their goals." In a later volume, however, McDougall eventually accepted the terms 'introvert' and 'extravert' as suitable for this factor, and as describing the temperamental types in which "emotional expression and action are inhibited or uninhibited" respectively; and he fully endorsed the view that, if all men were "ranged on a continuous scale from one extreme to the other, their distributions would be found to approximate to a normal curve" (2, p. 184).

³ E.g., *Grammar of Science* (1899), p. 436, Fig. 31. *Phil. Trans.* CC (1902), A, pp. 1-66. For the benefit of the non-mathematical student the principle on which the construction is based was explained as follows. Suppose the two hands of a clock represent two tests or traits. When the position of the two hands coincides (as at twelve o'clock), we may regard them as expressing a perfect agreement, i.e., a perfect positive correlation, between the two traits. When the hands point in opposite directions (as at half-past twelve), we may regard them as expressing complete antagonism, i.e., a perfect negative correlation. A position exactly midway, therefore (the two hands at right-angles,

by the cosine of the angle between a pair of radial vectors, and the strength and nature of each trait is indicated by the length and direction of the corresponding vector. If the vectors are taken to be of the same standard length, the cosine becomes identical with the product-moment correlation.

As applied to emotional impulses and conative tendencies generally, such a representation enables us to express, in a more explicit (though no doubt oversimplified) fashion, some of the fundamental principles of conational psychology. McDougall, it will be remembered, regarded these emotional tendencies as "mental forces,"¹ which set the ends, and sustain the course, of human activity"; and held that the more composite emotions could be resolved into more elementary tendencies, and that the elementary tendencies could be compounded to form the more composite. But if conational impulses can be regarded as forces, then it seems natural to enquire whether their mode of resolution and composition might not be effected in accordance with the familiar *parallelogram of forces*. This would imply that mental tendencies could be treated (at any rate to a first approximation) as obeying an *additive law*, though, of course, the mode of addition would be what is sometimes called geometrical addition, not arithmetical addition. Accordingly, as I argued in my paper, "if such an assumption could be verified empirically, it would yield a comparatively simple method of dealing mathematically with the problem of mental analysis and synthesis."

Thus, if each of the correlated traits we are assessing is regarded as a measurable force having a definite direction, we can, with the foregoing method of representation, proceed to resolve it into two hypothetical and uncorrelated forces, represented by vectors at right-angles to one another.² Further, if the correlations were such that the entire system of emotional tendencies with which we are concerned could be represented by a set of coplanar variables, then all the various traits in that system could be expressed in terms of the *same* two hypothetical forces, and could thus be represented on a two-dimensional diagram. In that case we could 'resolve' all the observed emotional tendencies into two independent 'components' of which they

e.g., one at XII and the other at III or IX) will express *zero* correlation. An angle less than a right-angle will express a positive but imperfect correlation: hands at 5 past XII would represent a correlation of 0.86, at 10 past XII, 0.50: so that, roughly speaking, the distances of arc between any two adjacent trait-points on the circumference of the circle will represent inversely the amount of positive correlation. Diameters, and chords drawn across the circle between two nearly opposite trait-points, will indicate the chief negative correlations. Thus we can use the horizontal diameter and the vertical diameter at right-angles to it to represent two completely uncorrelated bipolar factors; and, provided the correlations in which we are interested are due solely and exclusively to these two factors, the relations between all the traits and the factors can be represented on the diagram. In view of this explanation, such circular diagrams were referred to as 'clock diagrams' (cf. also *Distribution of Educational Abilities*, 1917, pp. 59f.).

¹ In suggesting that psychological analysis should aim at analysing tendencies or forces into their dynamic components, rather than sensory contents into their atomistic elements, McDougall was greatly influenced by the dynamic hypotheses of Fries, Beneke, and (above all) Lipps, as these in their turn had been influenced by the growing preference of the other natural sciences to deal with force or energy rather than with matter.

At first the above conception of psychological 'forces' may seem similar to that of the "directed forces, determinative of behaviour" which play so large a part in the 'dynamic psychology' of Lewin and his school. The chief difference would appear to be that the 'forces' postulated by Lewin, J. F. Brown, and others are expressly 'non-metricized' (on the difference see also McDougall, I, 24th ed., 1942, pp. 432, 435).

² Much of the following paragraphs will already be familiar to statistical psychologists in this country; but, as one object of this journal is to give "comparatively elementary and at times purely explanatory" accounts, it seemed that this simple enquiry would provide a useful opportunity for such an explanation. It may also perhaps help to remove some apparent misconceptions about my own attitude towards 'configurational diagrams' (e.g., as implied in Thurstone's criticisms, *Multiple Factor Analysis*, p. ix).

are the composite resultants ; and the resolution and composition could be effected in accordance with the parallelogram law.¹ Evidently the principle may, in theory at any rate, be generalized to include the case where the number of factors requires three or more dimensions.

To obtain such a representation in its simplest form, we can proceed as follows. Since in studying temperamental types we are not interested in the general factor, we must first eliminate its effects by the method of partial correlation, and then factorize the partial coefficients.² The partial coefficients thus obtained do not form a hierarchical table (5, p. 696 and Table II, *loc. cit.*) : hence they cannot be explained in terms of a single factor, and consequently will require at least two dimensions for their representation. If we assume that two factors alone suffice, then the square-sum of their saturations should amount to unity for every trait. That, however, is not the case. Nevertheless, the remaining factors include no significant common factor, and are for the most part comparatively small. Hence we may provisionally treat them as merely indicating 'errors of measurement,' i.e., irrelevant disturbances whose essential effect is simply to increase the unreliability of the trait-measurements, and so to reduce the size of the correlations obtained. Their influence can accordingly be eliminated by the usual formula for 'correcting for attenuation.' The net effect of this double adjustment will be to normalize the saturations for every trait, i.e., to increase each pair until their square-sum is unity, while preserving their original proportions or ratios. The augmented saturations obtained in this way with the present data are appended in the last two columns of Table II.

With these figures we can now readily construct a simple graph to represent the relations of the traits. We need two rectangular axes to represent the two bipolar factors ; each pair of augmented saturations can then be taken as specifying the Cartesian co-ordinates for the corresponding trait-point as referred to these two axes. Let us take the saturation for the first bipolar factor as measuring distance along the horizontal axis, and the saturation for the second bipolar factor as measuring distance along the vertical axis. On plotting the several points we shall then find that (in virtue of the normalization) each of them falls on the circumference of a circle with unit radius.

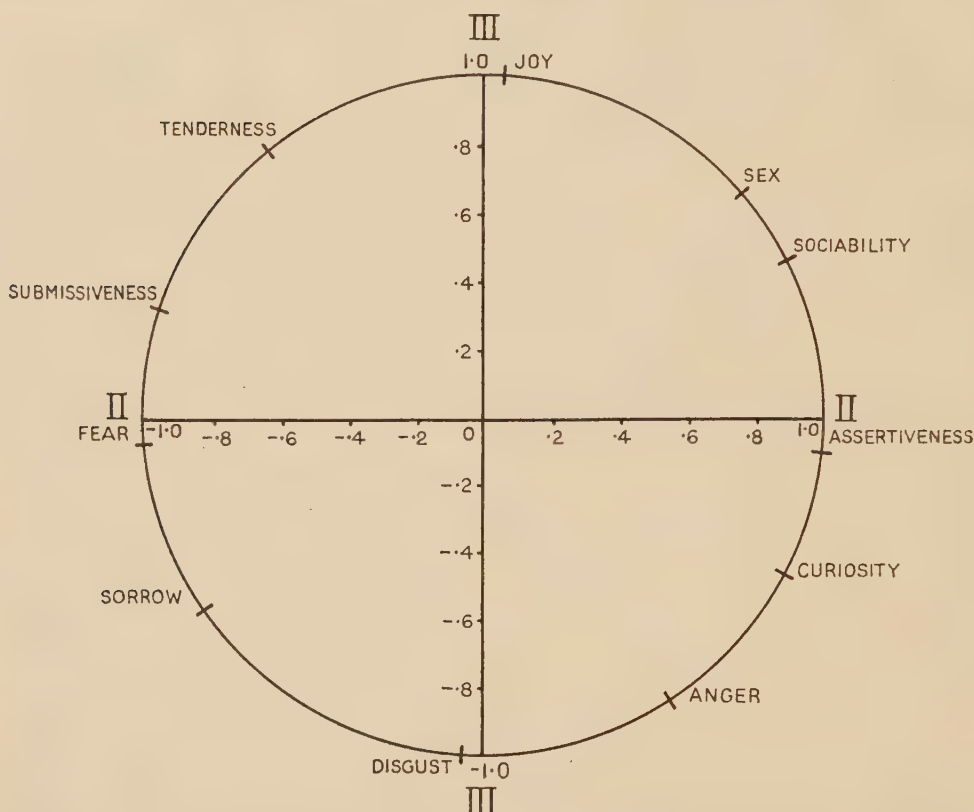
The result is illustrated in Fig. 2. In Fig. 1 I reproduce for comparison the corresponding diagram obtained by the same procedure in the previous investigation with a somewhat smaller group, and printed in my 1915 report (5, p. 696). It will be seen that the general arrangement displays a very close agreement—closer, indeed, than might have been anticipated from a mere comparison of the two tables of correlations or saturations.

The diagrams bring into clear relief several unexpected peculiarities. First, the specific correlations between every pair of adjacent traits are positive (since the distances of arc are in every case less than 90°). Thus the series of traits forms, as it were, a closed circle. The nearness of the several traits indicates which resemble each other most, and, roughly speaking,

¹ To forestall what I should regard as an irrelevant line of criticism, let me repeat what I said in my earlier paper, namely, that "these suggestions must be regarded merely as a description of a working hypothesis, not as a literal description of actual facts." Quoting Pearson's insistence on the limitations of the law even in the material universe (*Grammar of Science*, p. 319), I added that "we may have to supplement it by what has been called the 'hypothesis of modified action' ; we may even find that two emotional tendencies, instead of combining, may interfere with, alternate with, or even annul each other, so that the upshot may be something that was in no way deducible from our initial data." I ought also to say how much I was indebted, in endeavouring to formulate these dynamical ideas, to my colleague, Dr. Gwilym Owen (later Professor of Physics at Liverpool).

² The formula cited above (p. 181, footnote 1) suggests that we should seek the latent roots and vectors, not of R_c , but of $D^{-1} R_c D^{-1}$ —the reduced correlation matrix, corrected for 'attenuation' due to specific factors : the resulting saturations will not be precisely the same (see *Brit. J. Educ. Psych.*, VII, p. 193). As a practical working procedure I proposed applying the summation formula, not to the simple residuals, but to the partial correlations. Further corrections could then be applied, if necessary, at a later stage.

FIG. 1.—CORRELATIONS BETWEEN PRIMARY EMOTIONS
after elimination of general factor and correction for specificity
(1915 investigation)



II—II (horizontal axis) indicates first bipolar factor (sthenic-asthenic or extravertive-*versus*-introvertive: sthenic towards the right). III—III (vertical axis) indicates second bipolar factor (euthymic-dysthymic or pleasurable-*versus*-unpleasurable: euthymic towards the top).

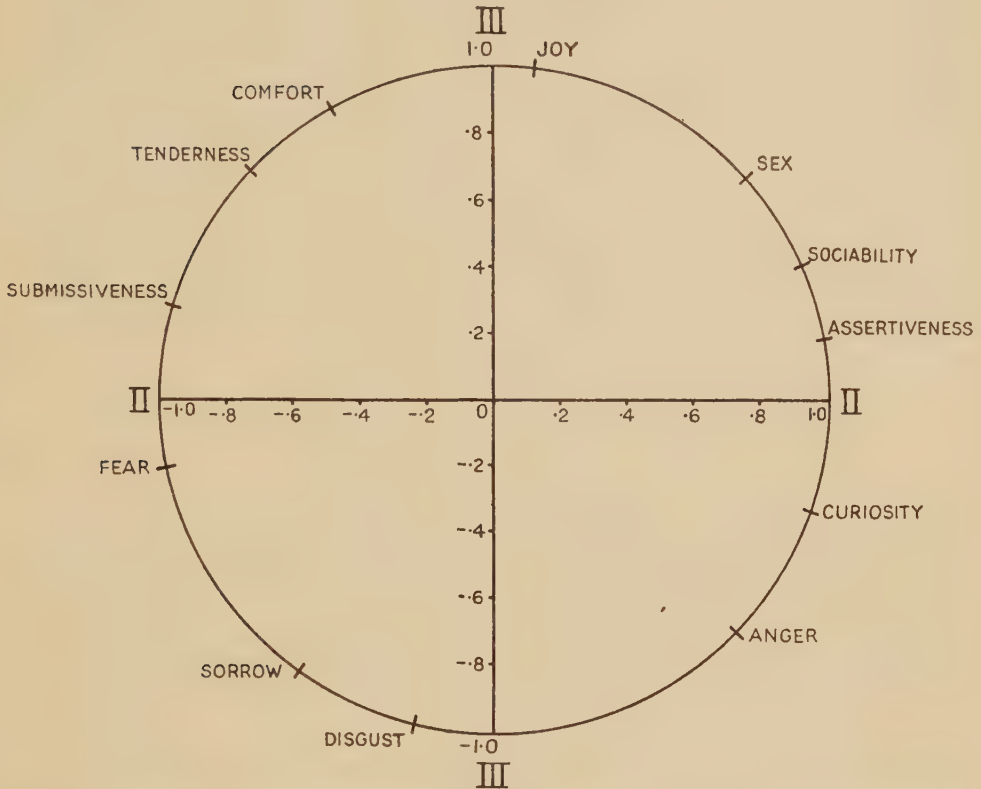
the amount of that resemblance. If we imagine chords drawn across the circle, passing through or near to the centre, we shall also perceive which tendencies have the largest negative correlations, and-so are mutually antagonistic.¹

A second corollary follows. In considering how to resolve the observed emotional tendencies, it is evident that we are not compelled to retain the original bipolar factors that emerged from the preliminary calculation. We may, if we wish, choose any other pair of orthogonal axes that happens to suit our *a priori* theories.² It is (in my view) this freedom

¹ McDougall's view of emotional inhibition was based on Sherrington's doctrine of the reciprocal inhibition of antagonistic muscles (or rather of the neural centres for those muscle-groups). The results of the analysis give his doctrine a clearer and more detailed interpretation, but at the same time, I think, indicate that the antagonism is rarely so complete as the analogy from reflexes might lead us to assume.

² With younger children (as contrasted with older) and with girls (as contrasted with boys), although the order and arrangement of the traits remain much the same, the factor-axes undergo a rotation clockwise from N to NNE and from E to ESE. The effect of this is that with them Joy appears as an asthenic (rather than sthenic) emotion: Disgust becomes sthenic, Assertiveness becomes more pleasurable, and Fear and Sorrow more unpleasurable.

FIG. 2.—CORRELATIONS BETWEEN PRIMARY EMOTIONS
after elimination of general factor and correction for specificity
(present investigation)

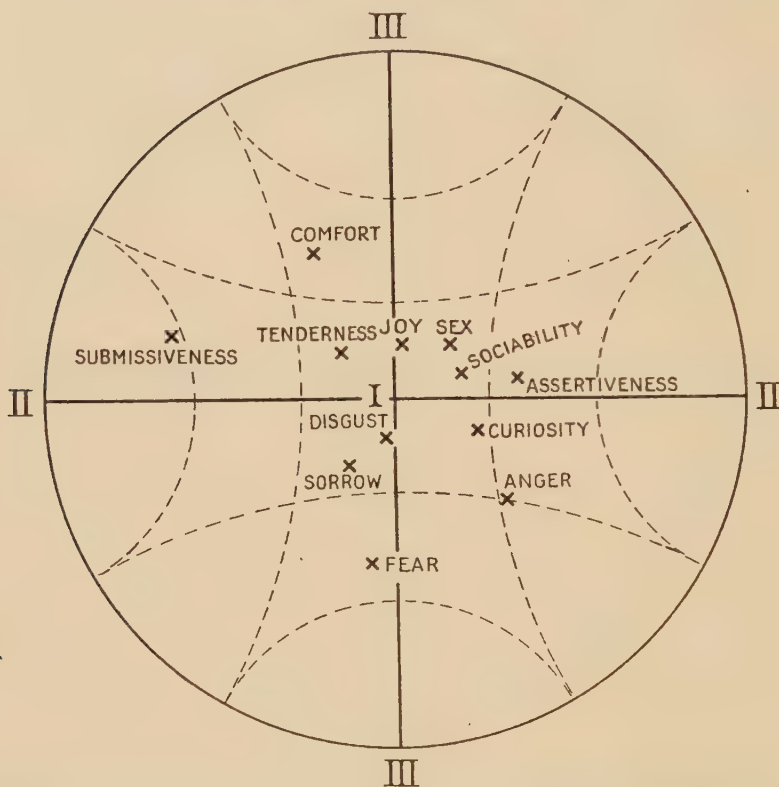


II—II (horizontal axis) indicates first bipolar factor (sthenic-asthenic or extravertive-*versus*-introvertive: sthenic towards the right). III—III (vertical axis) indicates second bipolar factor (euthymic-dysthymic or pleasurable-*versus*-unpleasurable: euthymic towards the top).

of choice that accounts for the number of slightly different schemes of temperamental types. As I pointed out in my original paper, and as Stern had also observed (*Differentielle Psychologie*, pp. 188, 481-5), nearly all these schemes turn on a fourfold classification, derived from two orthogonal bipolar components; but we now see that the directions of the paired components may differ appreciably from one writer to another, without affecting the relations between the traits themselves.

The Stereographic Net. We have, however, noted that several of the more recent schemes for classifying emotional tendencies require a three-dimensional representation. For this and other reasons it would be of interest to include in our geometrical model some representation of the first factor, as well as the other two. The most natural device would be to construct a kind of *Gefühlskugel* similar to Runge's well-known *Farbenkugel* as depicted in the older psychological textbooks (e.g., Wundt, *loc. cit. sup.*, II, p. 167, Fig. 191). To procure a representation on paper, i.e., in a single plane, the methods of projection adopted by geographers and others are (as I have shown elsewhere) extremely convenient. Various alternatives are available—globular, gnomonic, and stereographic. Stereographic nets are published

FIG. 3.—NORMAL CASES
STEREOGRAPHIC REPRESENTATION OF FACTORS AND TRAITS
Based on Correlations between Traits



I (centre) indicates the general factor ('general emotionality'); II—II (horizontal axis) indicates 'sthenic-asthenic' factor (sthenic towards the right); III—III (vertical axis) indicates 'euthymic-dythymic' factor (euthymic towards the top).

(Note that the method of plotting expresses the saturations as *proportions* of the common factor variance. Thus, judged by the absolute values in Table II, 'sorrow' has a *larger* general factor saturation than 'disgust': but, compared with its saturations for the bipolar factors (asthenic and dysthymic), its general factor saturation appears *smaller* relatively to 'disgust'.)

for the use of crystallographers¹; and these are particularly serviceable for exhibiting the relations between a number of traits whenever (as so frequently happens) we are concerned with three main independent factors only. The threefold factorial system revealed by the present investigation is shown in Fig. 3.

¹ For an elementary account of the commoner methods of projection, see A. R. Hinks, *Map Projections*. The nets, printed privately before the war, can now be obtained from Allen & Unwin, Ltd., 40, Museum Street, W.C.1; a boxwood 'stereographic scale' (purchasable from C. F. Casella & Co., Fitzroy Square, W.1) forms a useful adjunct. No more than a rudimentary knowledge of spherical trigonometry is needed to apply these devices. On the application of such methods to statistical problems, see *Factors of the Mind*, p. 81 and refs.

The traits are plotted in the same way as before, except that the saturations for the first *three* factors are normalized, instead of only the two main bipolars. This preliminary procedure may be regarded as a 'correction for attenuation due to specific factors.' If, on the three-dimensional sphere, the North Pole is taken to represent the general factor, then all the trait-points will fall on the Northern hemisphere, because their general factor-saturations are positive. The stereographic diagram may be regarded as showing this hemisphere as seen in perspective (i.e., projected on the equatorial plane or some plane parallel to it) with the spectator's eye at the South Pole.¹ The centre of the net will thus represent the general factor; and proximity to this centre may be taken as indicating the relative ² size of the general factor-saturation. The fact that most of the points are concentrated near the centre, and indeed can, with three minor exceptions, be enclosed with a right-angled spherical triangle (which would have its mid-point not far from the centre of the net), is due to the fact that, apart from these three exceptions, all the observed correlations between the traits are positive. It is one of the advantages of a stereographic projection, as contrasted with other methods, that it reveals such features very clearly, and enables suitable rotations or other transformations to be discerned by simple inspection, and trial-estimations procured by the usual methods for the graphical solution of spherical triangles.³

III. PSYCHONEUROTIC CASES

Factor Analysis of Correlations between Traits. We have already remarked how many investigators, who doubt whether anything approaching a classification into temperamental types can be verified among normal persons, nevertheless hold that such a classification is clearly discernible among the psychoneurotic. Others have contended that the classification thus observed among the neurotic cases is merely an exaggerated form of an identical classification found among the normal. To determine how far the two classifications agree, therefore, let us now apply the same factorial procedure to the data obtained from the psychoneurotic cases.

The group of cases with which we are concerned, and the symptoms which they exhibit, have been described more fully elsewhere (9, pp. 14f.). Since what are termed functional nervous disorders or 'psychoneuroses' are commonly considered to be primarily emotional or affective disorders, we may again restrict our analysis to emotional characteristics. The correlations between these traits are shown in Table III.

¹ To interpret the projection in detail, the reader should bear in mind the characteristics which remain invariant under this form of projection: (i) all circles, 'great' or 'small,' on the sphere are projected as circles on the stereographic plane (this includes straight lines which are circles with their centres at infinity); (ii) the angle at which two 'great circles' intersect on the sphere is projected into an angle of the same size on the stereographic plane. See Laboratory Notes on *The Use of Projection Methods*.

² That is, relative to the saturations of the same trait with other factors. To compare the size of the first factor-saturations for different traits, it is better to restrict the augmentation to the other two factors.

³ It may be added that the effects of selection, and the results of Pearson's trigonometrical correction-formulae, can also be very conveniently estimated in this way. But, as I have argued elsewhere, I consider such procedures useful as illustrative devices, but not as substitutes for explicit computation (cf. *Nature*, CLVI, 1945, p. 338). In other fields (e.g., for the solution of right-angled triangles in navigation and astronomy) experienced workers claim great accuracy (see Penfield, *Amer. J. Sci.*, XI, p. 25). Fedorov, however, claims that much higher accuracy is obtainable with gnomonic projection (*New Geometry*, p. 108). Most stereographic scales are also marked for use in gnomonic projections. For beginners globular projection is the easiest to understand and construct.

TABLE III. PSYCHONEUROTIC CASES: CORRELATIONS BETWEEN TRAITS

Trait	1	2	3	4	5	6	7	8	9	10	11	12
	Joy	Sex	Sociability	Assertiveness	Anger	Curiosity	Disgust	Sorrow	Fear	Submissiveness	Tenderness	Comfort
1. Joy	(.223)	.315	.161	.171	.156	.159	.095	.146	.256	.074	.179	.159
2. Sex	.315	(.656)	.402	.289	.407	.204	.088	.140	.279	.126	.147	.076
3. Sociability	.161	.402	(.264)	.143	.265	.111	.040	.053	.117	-.038	.061	.041
4. Assertiveness	.171	.289	.143	(.437)	.501	.294	.166	-.043	-.123	.059	-.044	.171
5. Anger	.156	.407	.265	.501	(.763)	.352	.290	.010	-.131	.139	-.044	.303
6. Curiosity	.159	.204	.111	.294	.352	(.264)	.205	.076	.041	.132	.081	.222
7. Disgust	.095	.088	.040	.166	.290	.205	(.314)	.263	.240	.243	.209	.353
8. Sorrow	.146	.140	.053	-.043	.010	.076	.263	(.469)	.619	.189	.401	.393
9. Fear	.256	.279	.117	-.123	-.131	.041	.240	.619	(.924)	.307	.582	.337
10. Submissiveness	.074	.076	-.038	.059	.139	.132	.243	.189	.307	(.247)	.259	.338
11. Tenderness	.179	.147	.061	-.044	-.044	.081	.209	.401	.582	.259	(.448)	.365
12. Comfort	.159	.126	.041	.171	.303	.222	.353	.393	.337	.338	.365	(.503)

The correlations have been factorized by the same method as before ; and the results are given in Table IV. Four factors have been extracted. But once again the last is devoid of all statistical significance.

TABLE IV. PSYCHONEUROTIC CASES : FACTOR-SATURATIONS

Trait	I	II	III	IV	Sum of Squares
1. Joy	·377	·114	·185	·168	·217
2. Sex	·563	·420	·387	—·061	·647
3. Sociability	·309	·279	·231	—·107	·238
4. Assertiveness	·363	·454	—·243	·104	·408
5. Anger	·542	·533	—·365	—·219	·759
6. Curiosity	·385	·203	—·195	·115	·241
7. Disgust	·451	—·131	—·276	—·089	·305
8. Sorrow	·489	—·429	·121	—·125	·454
9. Fear	·621	—·568	·445	—·028	·907
10. Submissiveness	·364	—·242	—·204	·050	·235
11. Tenderness	·476	—·415	·146	·153	·442
12. Comfort	·614	—·218	—·232	·039	·480
Sum of Squares	2·693	1·601	·873	·167	5·333
Contribution to Variance ..	22·4%	13·3%	7·3%	1·4%	44·4%
Physical Symptoms. . .	·389	—·267	—·328	—	—

The factor-pattern exhibits certain limited resemblances to that which we have obtained from normal children. The first factor is a general factor ; the second a bipolar factor distinguishing sthenic or extravertive traits from asthenic or introvertive ; the third factor shows some similarity to the previous fourth ; and the fourth to the previous third.

But the differences between the two factor-patterns appear far more instructive than the resemblances. To begin with, the general factor now accounts for only 22 per cent. of the total variance instead of 37 per cent. ; and the relative size of the factor-saturations shows a different order. These features could readily be accounted for as the effects of selection. Elsewhere I have shown that the most conspicuous feature distinguishing the psychoneurotic child from the normal is his increased emotional instability. Evidently, therefore, in selecting psychoneurotic cases, we have been taking cases from the upper end of the scale of general emotionality. The effect of this must be to curtail the range of variation for the general factor. Hence its saturation coefficients are automatically diminished.

A number of investigators who have carried out factorial studies on neurotic cases have found a similar general factor. Many have inferred that, since this factor is common to all the traits assessed for such groups, it can therefore be designated a "general factor of neuroticism" (cf. 16, pp. 41, 257 and refs.). But a general factor for neuroticism would be a factor that is found among neurotic cases only or among traits peculiar to neurotics ; and

to prove such a conclusion it would be essential to examine a control group of normal persons, and demonstrate that the factor in question is not found among the normal.¹

The second factor contributes a far larger proportion to the common factor variance than before. It seems plain that, as a result of selection, we are now dealing with a population in which the contrast between 'extraverts' (or 'cyclothymes') and 'introverts' (or 'schizothymes') has become much sharper than in the normal population. There is still a considerable overlap; and it would be impossible to allocate every case decisively to one type or to the other. However, a corresponding factor analysis of adult cases, and of older children as compared with younger, shows that, as the groups grow older, the overlap becomes smaller, and the distinction becomes increasingly clear.

The third factor is of a somewhat novel kind. Its nature is perhaps best determined by examining those individuals who show large positive or negative factor-measurements for it. We then discover that it divides both the extraverts and the introverts into two subclasses. The extraverts are divided into (a) those characterized by strong sexual propensities, joyful moods, and a fondness for company, and (b) those characterized by aggressive, assertive, and inquisitive tendencies. Nearly all who have been diagnosed as hysterical fall into the first subdivision, and all the obsessional and compulsive cases into the second. With the first the extraversion takes the form of a cheerful exhibitionism: they love to be always in the social limelight, and show an excessive interest in the opposite sex; with the second it takes the form of an obstinate self-will, a desire to resist or domineer.² The introverts are similarly divided into (a) those who are characterized by fear or anxiety, together with a marked degree of affection and a proneness to grief, and (b) those who are characterized by disgust,³ submissiveness, and a desire for comfort. Nearly all the cases of anxiety-neuroses fall into the former group, and the majority of the neurasthenic cases into the latter. An earlier analysis of adult cases has shown that physical symptoms, both organic and functional, were far commoner among the cases of hysteria and neurasthenia (9); and accordingly I have calculated the correlation between the various factor-measurements and the frequency of physical symptoms. The coefficients⁴ are appended at the foot of Table IV. It will be seen that physical

¹ One might just as well argue that, because the possession of a backbone is the most conspicuous feature of the mammalian body, therefore mammals may be defined as creatures possessing backbones. The fallacy here observable is by no means uncommon in factorial studies. As I have argued elsewhere, "to define the specific nature of a general factor, that factor must first be turned into a group factor." Of course, if we take the view that functional nervous illnesses are not really illnesses at all, but merely extreme manifestations of the general emotionality found in all persons, then the argument ceases to be fallacious; but in that case, to name so universal a quality a 'factor of neuroticism' would seem somewhat misleading. However, the popular notion of a general 'neuropathic diathesis' certainly deserves fuller investigation by statistical techniques.

² It would seem that the factor-pattern changes somewhat with age. In a pre-war study of the case-histories then available, Miss J. F. Pemberton reported that "the hysterical group seems characterized by 'positive self-feeling' (assertiveness); the neurasthenic by 'negative self-feeling' (renunciation or submissiveness); the compulsives by pugnacity; and the anxiety-neuroses by fear." Her cases were fewer; and their ages for the most part between 6 and 12. She noted, however, that "with increasing years the emotional pattern becomes more complex: towards puberty the hysterical group displays stronger sex and social propensities; they also show less assertiveness and more suggestibility, while the compulsives show increasing contra-suggestibility." "It might be better," she concludes, "in the case of younger children to change the labels, and speak of 'assertive neuroses,' 'anger neuroses,' 'fear neuroses,' and 'submissive neuroses.'" (Better still, I would add, to drop the term 'neuroses.') At the same time it should be noted that the factor-pattern, whether simple or complex, merely professes to depict the type: the special characteristics of each particular case vary widely with the past history and the present circumstances of the individual child.

³ On the relation of disgust to fatigue and other symptoms distinctive of neurasthenia see 10, p. 229.

⁴ The formula used is that given in *Journ. Psych.*, VI, 1938, p. 355, eq. vii. For the physical symptoms chiefly noted in the several groups see 9, p. 15.

symptoms are associated more closely with introvertive tendencies than with extra-vertive, and (if we may trust the group factor analysis) with the hysterical and the neurasthenic subgroups than with the subgroups marked by obsessions, compulsions, or anxiety-states.

The previous third factor, which distinguished pleasurable from unpleasurable emotions, has now almost vanished. Traces of it may perhaps be discerned in the fourth factor ; but this factor, as we have seen, is not statistically significant.

The conclusions suggested by the whole analysis may be summed up as follows. Among children of school age, who have been provisionally described or diagnosed as 'psychoneurotic,' at least four common syndromes, and four corresponding groups of patients, are distinguishable. These appear roughly parallel to the conditions described as 'hysteria' (with conversion hysteria as the typical form), 'anxiety-states,' 'obsessive-compulsion states,' and 'neurasthenia' respectively.¹

A rigorous investigation would require a test of significance for the distinctions thus drawn. Unfortunately significance-tests for factorial results are still in an unsatisfactory state. Applying the chi-squared test in the form in which I originally proposed it, we find that the first two bipolar factors are fully significant. An alternative test can be obtained by employing Bartlett's chi-squared approximation for the likelihood ratio. Here the natural procedure would be to calculate factor-measurements for each type. The bipolar factors are first converted to group factors ; but, since approximate weights are sufficient, the computation is not unduly lengthy. The determinantal criterion can then be applied to the product-sum matrices derived from these factor-measurements : using the approximate form of the test, we again find that the 'variance between groups' is fully significant. The smallest group-difference is that between the neurasthenic and anxiety groups : for this the critical ratio is 3.7. It is thus fully significant ; and we may consequently infer, *a fortiori*, that the differences between all other pairs are significant.²

In the general literature on psychoneurotic conditions among school children the most recent account is that given by Dr. Barton Hall in her chapter on that subject (18, Ch. X). Up to a point our two classifications correspond. The chief difference is that Dr. Hall is disinclined to accept any separate group or disorder designated 'neurasthenia'—one of the terms, she says, that "should no longer be employed, so far as children's ailments are concerned." My data, however, seem definitely to indicate a distinctive syndrome, closely agreeing with that so named by Prof. Aubrey Lewis³ (17, pp. 1816-17). At the same time, it should be emphasized that the whole analysis claims to offer no more than a provisional classification of 'surface-traits' into 'phenotypes' (if I may borrow Dr. Cattell's phraseology), and not a classification of 'source traits' into 'causal types.' Further, the classification is based merely on *tendencies* towards a typical pattern ; so that the various groups display a considerable degree of overlapping.

With children, as I have pointed out elsewhere (10), a large proportion—probably the majority—of those diagnosed as cases of 'mental illness' are not really 'ill,' in the sense that they are suffering from some pathological disorder that is qualitatively distinct from any normal deviation in normal characteristics, however exceptional the deviation may be. Thus, on the affective side the position is much the same as it is on the cognitive side. The so-called mentally defective include numerous cases that are characterized merely by an

¹ A fuller account of these conditions as met with among school children, together with figures for the frequencies of each type, will be found in my book on *The Subnormal Mind* (1937), Ch. VI, 'Asthenic Neuroses,' and Ch. VII, 'Sthenic Neuroses.'

² The general procedure was more fully described in a War Office Memorandum on 'Allocation of Recruits on the Basis of Test Results.' For formulæ see 13, pp. 336-41, 362 and refs.

³ His description (omitting one or two items less frequent in children than in adults) admirably fits a group of my cases, numbering about 12 per cent. of the total (rather less perhaps among the girls) : "a form of irritable hypersensitive weakness and depression, not uncommon after infections, exhausting experiences (hunger, worry, hæmorrhage), cranial injuries, and chronic poisoning. The symptoms are partly somatic . . . On the psychological side there are feelings of languor, incapacity to concentrate on mental work, doubts over accuracy of memory, loss of interest, lessened control of emotions, . . . hypochondriac trends" (Lewis, *loc. cit. sup.*, p. 1816).

exceptional but normal deviation below the general average, together with a smaller number who differ qualitatively from the normal and consist of definitely pathological cases, classifiable into various subgroups. With cognitive abilities the difference is partly a genetic one: the former are cases of multifactorial inheritance, the latter chiefly of unifactorial inheritance or of early post-natal or pre-natal disease. How far this genetic difference holds good of the psychoneurotic cases is too uncertain a problem to be dealt with here.¹

Correlations between Persons. On one important question, however, a statistical analysis may claim to throw some light. During the last 35 years, since the work of the psychoanalytic writers and the experiences of the 1914-18 war first directed attention to psychoneurotic disorders, there have been prolonged controversies as to whether psychoneurosis is a single condition varying solely in degree, or whether it covers a heterogeneous collection of essentially different illnesses. My own view is that psychoneurotic disorders form a fairly distinct genus, but that within that genus there are also fairly distinct species. In psychological medicine, however, all attempts at classification meet with considerable difficulty, because, whatever may be the case with purely physical disorders, mental disorders (at any rate in the present state of knowledge) cannot be grouped into sharply demarcated 'clinical entities,' after the fashion of the 'Linnæan classifiers' of the early nineteenth century. The mixed, the borderline, and the transitional cases are far more frequent than the typical cases of the textbook: normal types of personality overlap with the various neurotic types, the neurotic types with the psychotic, to a degree seldom appreciated in the discussions of differential diagnosis. It is, indeed, precisely for this reason that I have argued that, until further research has furnished us with more accurate knowledge, the only sound basis for a working classification is that supplied by factor analysis.

Our ultimate aim is to classify persons or patients rather than symptoms or traits. Now, when we are concerned with the classification of persons, what we have to compare are the factor-measurements for persons, not the factor-saturations for traits. And, in virtue of the 'reciprocity principle,' so long as we are concerned with the bipolar factors only, we can obtain these factor-measurements much more readily, and in my view much more appropriately, by factorizing correlations between persons²: for the crucial question now is not how far does this *trait* (or group of traits) resemble that, but how far does this *person* (or group of persons) resemble that.

If then we take the 40 symptomatic traits which appear most relevant in the differential diagnosis of psychoneurotic disorders (9, p. 15), we can mark each of them for its presence, absence, or degree of severity in every person. We can then correlate these persons, and so determine the chief groups or clusters into which they fall. In a rigorously complete enquiry, we should proceed to demonstrate the validity of such groupings by applying the usual tests of statistical significance, as laid down for

¹ I believe that general emotional instability is to a large extent an inborn quality, attributable to multifactorial inheritance; but there is some small degree of evidence to suggest that differences in temperamental type may, in certain instances, be due to unifactorial inheritance, and that the grading is due mainly to non-genetic conditions (in part no doubt endocrinological); most cases of psychoneurotic illness, however, are precipitated by the double operation of innate constitution and environmental stress or maladjustment (cf. 10, pp. 205f.; also *Eugenics Rev.*, IV., pp. 22f.).

² As a method of distinguishing groups or types, the technique of correlating persons has been in use since 1912 (cf. Burt, C., *Brit. J. Psych.*, XXVIII, p. 59 and refs.). For its application to temperamental assessments, with a demonstration of the 'reciprocity principle,' see 9, p. 179 and refs. My former colleague, Dr. Stephenson, has questioned the validity of the principle (for our respective views see the joint article in *Psychometrika*, IV, pp. 269-282). However, even if the reciprocity principle is denied, the conclusions here drawn about the relations of the groups would remain unaffected.

a multivariate analysis of this nature. Once we have satisfied ourselves of the existence of such groups, we can treat each group as a composite person. We can then mark or grade the various traits in order according to their intensity or frequency, and correlate the groups. The factor-saturations obtained from these group correlations will thus reveal the relations of the several groups to each other.

TABLE V. PSYCHONEUROTIC CASES : CORRELATIONS BETWEEN GROUPS

Group	Normal	Anxiety-states	Neurasthenia	Hysteria	Obsessions	Psychopathic
Normal	(.726)	.599	.641	.486	.201	.070
Anxiety-states	.599	(.869)	.664	.071	.146	-.073
Neurasthenia	.641	.664	(.784)	.253	.008	-.136
Hysteria	.486	.071	.253	(.635)	.206	.212
Obsessions	.201	.146	.008	.206	(.211)	.170
Psychopathic	.070	-.073	-.136	.212	.170	(.184)

Unfortunately the data available for an elaborate enquiry on these lines is not very accurate nor yet very extensive. Those who conduct the case-examinations are occupied more with the immediate problem of treating, and if possible curing, their patients ; and time, as a rule, forbids the systematic collection of a complete case-study for every child, with a view to scientific analysis later on.¹ Accordingly, the results here described can offer no more than a rough and tentative survey.

TABLE VI. PSYCHONEUROTIC CASES : FACTOR-SATURATIONS

Group	I	II	III	Sum of Squares
Normal	.842	-.067	.109	.726
Anxiety-states	.704	-.541	-.283	.869
Neurasthenia	.685	-.484	.283	.784
Hysteria	.576	.465	.294	.635
Obsessions	.291	.244	-.258	.211
Psychopathic	.132	.383	-.145	.184
Sum of Squares	2.108	.954	.347	3.409
Contribution to Variance ..	35.1%	15.9%	5.8%	56.8%

¹ There is a most pressing need for such research ; and one of the objects of this article is to persuade those who work in child guidance clinics and centres to gather more extensive and more trustworthy material in accordance with a preconceived scheme. Several writers have quoted the statement made by a former Medical Director of the Child Guidance Council to the effect that " since child guidance became a branch of medicine it has become more scientific." But as a matter of fact the change from psychological to medical direction has been followed by a remarkable diminution in the number of researches dealing with child guidance problems in London, in part no doubt owing to war-time difficulties. At the moment there is a growing interest in this field of research ; yet few investigators seem to realize the need for planning such enquiries in advance. Only if the needs of the eventual statistical analysis are borne clearly in mind during the case-examinations—only, that is to say, if the material is carefully compiled according to a predetermined programme—are adequate and appropriate data likely to be obtained. The subsequent investigation of case-histories, which have been filed by different workers interested solely in practical problems, is likely to be as futile or misleading in the field of psychoneurosis as it was in the earlier enquiries by school medical officers and others, forty years ago, on problems of mental deficiency.

My own cases have been grouped into six main classes in accordance with the scheme suggested by the preliminary factorization of traits. This yields 97 cases of 'anxiety-states,' 73 of 'neurasthenia,' 46 of 'hysteria,' and 34 of 'obsessions or compulsions.' To these are added 183 'normal' and 28 described as 'psychotic, pre-psychotic, or psychopathic.' The reduction in the total number of normal and psychoneurotic cases is due to the fact that the information secured for the remainder was not sufficiently detailed for the method proposed. The correlations between the groups are set out in Table V. A table of correlations between six variables can always be expressed in terms of three uncorrelated factors. And the factor-saturations, calculated by simple summation, are shown in Table VI.

The general factor obtained by correlating persons is, of course, different from that obtained by correlating traits. Here it denotes the general emotional pattern to which all persons and all groups more or less conform. The group whose profile approximates most closely to this hypothetical pattern proves, naturally enough, to be the normal group. The others conform less closely, with varying degrees of divergence—the anxiety-states but little, the psychotic or psychopathic group (as we might expect) diverging most of all. The bipolar factors that are obtained by correlating persons should be essentially the same as those obtained by correlating traits. In Table VI the first bipolar factor evidently distinguishes the sthenic or extravertive psychoneuroses (hysteria and obsessional disorders) from the asthenic or introvertive (anxiety-states and neurasthenic conditions); the second factor appears to differentiate those conditions which are commonly marked by physical disturbances (conversion-hysteria and neurasthenia) from those in which physical symptoms are less conspicuous (anxiety-states and obsessional disorders). This is confirmed by a more detailed study of the frequency of physical symptoms: it is still more obvious in adult cases (9, p. 17: with adults, it may be added, and with many of the older children the anxiety-neuroses, as distinct from other anxiety-states, tend to get classed with the former group, i.e., that showing physical symptoms).

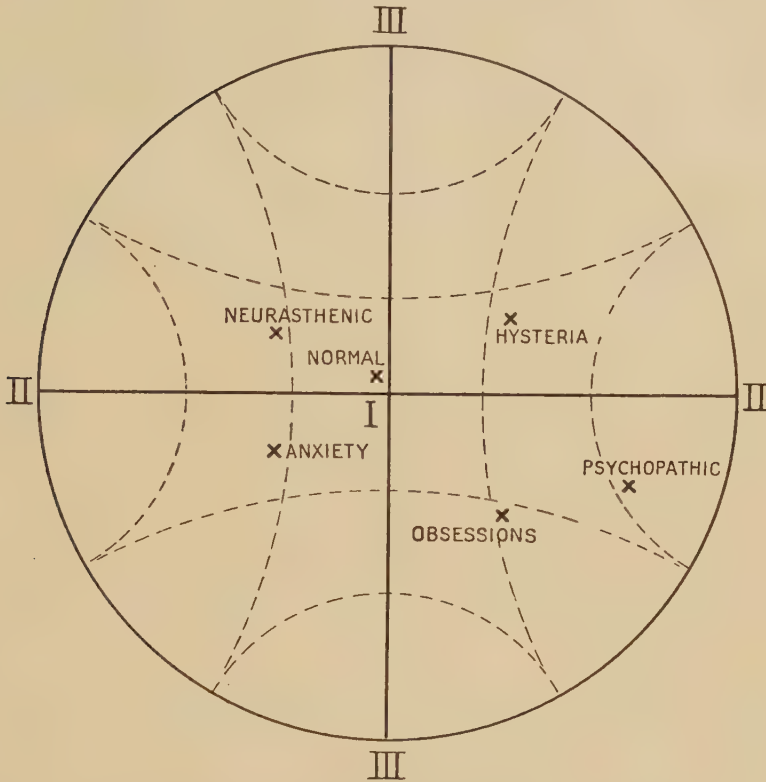
Stereographic Representation. Since only three factors are required to describe the various groups and their interrelations, it is possible to represent the results once again by a two-dimensional diagram based on a stereographic projection (Fig. 4). The proximity¹ of the points representing the groups shows which groups resemble each other most of all, that is to say, which are most likely to occasion difficulty when a differential diagnosis is under consideration.

As we have already seen, in view of the large amount of overlapping between different types of mental disorders (as contrasted with the relatively clear lines of demarcation supposed to obtain in the case of physical disorders), several writers have advanced the view that the fundamental difference is solely one of severity or degree, and that the alleged qualitative distinctions are by comparison small, incidental, and irrelevant. If, they argue, the normal merge into the neurotic and the neurotic into the psychotic by almost imperceptible gradations, that is because all are manifestations, with differing intensities, of one and the same type of reaction. Such a 'gradation theory' was favoured by Spearman and his school, who (quite rightly) always questioned the assumption of group factors unless their statistical significance had been conclusively proved. And it has recently been put forward by several psychiatrists both in this country and abroad.²

¹ As I pointed out in my original *Report*, there are many alternative methods of determining such proximity. I have discussed them more fully in a recent memorandum for the National Foundation of Educational Research on 'The Allocation of Children to Different Types of School' (a practical problem in which once again the essential difficulty is to discover a convenient but rigorous statistical procedure for discriminating between overlapping groups that differ in different qualitative directions and not merely in degree). The earliest of such methods, from which directly or indirectly most of the others derive, is Pearson's 'coefficient of racial likeness' (4). Of more recent suggestions one of the most ingenious is that proposed by Prof. P. Delaporte, 'Une méthode statistique de recherche des types' (*Biotypologie*, VIII, pp. 14-23).

² Cf. Curran, D., 'The differentiation of neuroses and manic-depressive psychoses,' *J. Ment. Sci.*, LXXXIII, 1937, pp. 156-174, and Bowlby, J., *Personality and Mental Illness*, 1940. The same hypothesis is implied by the practice adopted by many medical writers of attempting to clarify and describe normal personalities in terms drawn from mental pathology.

FIG. 4.—PSYCHONEUROTIC CASES
STEREOGRAPHIC REPRESENTATION OF FACTORS AND TRAITS
Based on Correlations between Persons (Grouped)



I (centre) indicates the general factor ('general emotional pattern'). II—II (horizontal axis) indicates the 'sthenic-asthenic' 'introvertive-versus-extravertive' factor (introvertive towards the left); III—III (vertical axis) indicates 'physical-versus-non-physical factor' (physical symptoms towards the top).

Were the gradation hypothesis true we should expect that the contribution made to the total variance by the bipolar factors would be negligible as compared with the contribution of the general factor. In that case the central points for all the groups would fall approximately on the same straight line. The graph shows at once that this is far from being so. The fact that the various groups are removed from the centre by different distances certainly indicates that they are in part distinguished by differing degrees of abnormality; but the fact that they are removed in different directions implies that in addition they differ qualitatively amongst themselves.¹

¹ In the *Report* already cited, two supplementary methods were also employed, and gave much the same results: viz. (i) the percentage of overlap between medians (cf. *J. Exp. Ped.* I, 1912, p. 283) and (ii) a procedure analogous to that of the 'discriminant function,' namely, taking the groups in pairs (beginning with the hardest to discriminate), and calculating the partial regressions for a biserial point-distribution ($w' = kd'R^{-1}$), and then using $w'd = kd'R^{-1}d$ (a coefficient not unlike Mahalanobis' 'generalized distance,' cf. 13, p. 359) to measure proximity. Such calculations can be carried out most systematically by applying the formula for a multiple discriminant function (cf. Burt, this

In determining to which of two or more classes a particular child belongs I suggest following the same principle as I proposed for determining whether he belonged to the normal or mentally defective group (11, p. 177). If we know the proportions of the two populations falling into either group, and the means and standard deviations for each in terms of the factor which forms the basis of the distinction between the two groups (the general cognitive factor in the case of mental defectives), then, assuming that both follow an approximately normal distribution, we can determine the likelihood that a child with a given deviation will fall into one group or the other. The borderline will therefore be given by a deviation for which the probability that he belongs to either group is exactly equal. This leads to a simple but serviceable formula : (cf. 13, pp. 271, 295f.).

In conclusion it should be emphasized that the results here described hold good only for alleged psychoneurotic cases among children. Among adults, as I have shown in a previous paper, the factorial results suggest that such disorders are more specialized, and more frequently pathological. This, however, is a problem which I hope will shortly form the subject of a joint contribution by two of my co-workers.

It remains for me to express my thanks to the teachers, research-students, and others who assisted in the collection of the data.

IV. SUMMARY AND CONCLUSIONS

1. The correlations between assessments for primary emotions, obtained from approximately 500 normal children and over 300 psychoneurotic cases, have been factorized. In both cases at least three significant factors are discernible. But the factors obtained from the psychoneurotic group are only partly identifiable with those obtained from the normal.

2. With the normal children the following factors are fully significant : (i) a general factor of emotionality ; (ii) a bipolar factor distinguishing sthenic (or extravertive) from asthenic (or introvertive) emotions ; and (iii) a further bipolar factor distinguishing euphoric (or pleasurable) from dysphoric (or unpleasurable) emotions.

3. The two bipolar factors yield a fourfold cross-classification not unlike that described in the traditional theory of 'temperamental types.' On plotting a two-dimensional diagram to show the pattern produced by these two factors, we obtain from this large group a scheme very similar to that obtained in a much earlier investigation from a smaller group by somewhat rougher procedures. The relations of the several traits to each other and to all three factors can be exhibited most clearly by using printed stereographic nets, which furnish a more complete representation of a three-dimensional configuration.

4. With the psychoneurotic group the first two factors appear to be much the same as those obtained from normal children, except that the effects of selection are revealed by alterations in the factor-variances and factor-saturations : the third

Journal, II, p. 105 and refs., and *Psychometrika*, IX, p. 225). The latter method was originally proposed for studying the problem of allocating pupils at 11 + to different types of school ; and where, at most, only 3 or 4 groups are concerned appears satisfactory, though cumbersome. Further experience with Army data, however, indicates that, with a larger number of groups, particularly when the groups depart considerably from a linear arrangement, the device is not very suitable. For most practical purposes, provided the traits or test-items can be made sufficiently numerous, the method of correlating persons seems to be at once the simplest and the most effective. For the more elaborate methods needed when significance has to be rigorously tested see 13, pp. 271, 345, and refs.

factor, however, differs markedly from that obtained with the normal group. The general factor appears to represent, not 'neuroticism' but 'general emotionality.' The two bipolar factors suggest a classification of neurotic cases into four overlapping groups—roughly identifiable with cases of anxiety-states, obsessional-states, hysteria, and neurasthenia respectively.

5. With the psychoneurotic cases the correlations between persons and groups of persons have also been factorized. The correlations between the six main groups can be accounted for by three factors. These enable the relations between the groups to be represented in a 'space' of three dimensions; and once again the method of stereographic projection supplies a convenient diagrammatic representation. The evidence suggests that the six groups—the normal, the four psychoneurotic, and the psychopathic—differ qualitatively as well as quantitatively.

6. Provided the term 'type' is taken to mean a pattern of factor-measurements to which individuals may approximate in varying degrees, and not one of two mutually exclusive classes, the analysis appears fully to confirm the existence of the most commonly cited pair of 'temperamental types,' namely, the 'objective,' 'tough,' or 'extraverted' and the 'subjective,' 'tender,' or 'introverted' respectively. The distinction is found both among the normal and among the psychoneurotic. Moreover, of all such classifications, this would appear to be the most important, in the sense that it accounts for the largest proportion of the individual variance. Nevertheless, it does not account for so large an amount as the factor of 'general emotionality.'

REFERENCES

1. McDougall, W. *Introduction to Social Psychology*. London: Methuen & Co., 1908.
2. McDougall, W. *The Energies of Men: A Study of the Fundamentals of Dynamic Psychology*. London: Methuen & Co., 1932.
3. Pearson, K. "On lines and planes of closest fit to systems of points in space." *Phil. Mag.*, II, 1901, 559.
4. Pearson, K. "On the coefficient of racial likeness." *Biometrika*, XVIII, 1926, 105.
5. Burt, C. "General and special factors underlying the primary emotions." *Brit. Assoc. Ann. Rep.*, LXXXV, 1915, 694.
6. Burt, C. "The mental differences between individuals." *Brit. Assoc. Ann. Rep.*, XLI, 1923, 215-39.
7. Burt, C. "The analysis of temperament." *Brit. J. Med. Psych.*, XVII, 1938, 158f.
8. Burt, C. "The factorial analysis of emotional traits." *Character and Personality*, VII, 1939, 238-54, 285-99.
9. Burt, C. "The assessment of personality." *Egypt. J. Psych.*, II, 1946, 1-21.
10. Burt, C. *The Subnormal Mind*. London: Oxford University Press, 1935.
11. Burt, C. *Mental and Scholastic Tests*. London: Staples Press, 1921.
12. Piéron, H. "La question des types psychophysiologiques." *Biotypologie*, VIII, 1946, 1-13.
13. Kendall, M. G. *The Advanced Theory of Statistics*, II. London: Griffin & Co., 1946.
14. Oates, D. W. "Group factors in temperament qualities." *Brit. J. Psych.*, XX, 1929, 118-136.
15. Cattell, R. B. *Description and Measurement of Personality*. New York: World Book Company, 1946.
16. Eysenck, H. J. *Dimensions of Personality*. London: Kegan Paul, Trench, Trubner & Co., 1947.
17. Lewis, A., and Mapother, E. Section on "Psychological Medicine," ap. Price, F. W. (ed.), *Textbook of the Practice of Medicine*. London: Oxford University Press, 1937.
18. Hall, M. Barton. *The Psychiatric Examination of the School Child*. London: Arnold and Co., 1947.

PRIMARY PERSONALITY FACTORS IN WOMEN : A RE-ANALYSIS

By CHARLOTTE BANKS

Department of Psychology, University College, London

I. *The Original Centroid Analysis.* II. *Re-analysis by Simple Summation : Methods Proposed (a) for Bipolar Analysis, (b) for Group Factor Analysis.* III. *Results Obtained. Interpretation of Factors : Comparison with (a) Unrotated and (b) Rotated Factor Matrix.* IV. *Summary and Conclusions.*

I. THE ORIGINAL CENTROID ANALYSIS

In the last issue of this *Journal* (1) Dr. Raymond Cattell published a table giving the correlations between the ratings for 36 traits as observed in 240 College women. The correlations were first factorized by Thurstone's centroid method (8, pp. 232f., 9, pp. 161f.). Twelve factors, all bipolar, were obtained. The centroid factors were then rotated graphically to obtain what is described as a 'simple structure' (9, pp. 194f.). For this 588 rotations were required, involving 18 months of full-time work, and the assistance of three computers. The final matrix contains eleven group factors of a bipolar type, which are discussed in detail and compared with corresponding factors previously obtained from men (2).

Results based on so large a collection of traits are naturally of great interest and importance. Nevertheless, the procedure adopted seems to have entailed an unnecessary amount of labour ; and even so the conclusions reached do not seem entirely satisfactory or convincing. To begin with, the correlations between the traits and the rotated factors appear unexpectedly small ; and, so far as can be judged, do not yield a very close fit to the original correlation matrix. Secondly, it is difficult to believe that as many as 11 or 12 significant centroid factors can be deduced, when the initial assessments have been obtained from only 240 persons, and when the majority of the observed correlations are no larger than 0.5.¹ Thirdly, the factors finally reached overlap enormously ; ten of their intercorrelations are over 0.35, and one is as high as .52. Evidently, by proceeding in this way, not one but a number of different and equally plausible factor-matrices could be secured. There is therefore no guarantee that the particular set of factors eventually reached in Table IV is in any way conclusive ; and, as we shall see in a moment, it possesses several inherent improbabilities of its own.

For these reasons, it has occurred to more than one reader that it would be instructive to attempt a fresh analysis of the data by one or other of the alternative techniques employed in this country. Accordingly, I have ventured to analyse the data afresh, and to apply Burt's methods of summational analysis (5, 7) with a view to obtaining, first, a set of general and bipolar factors (corresponding to Cattell's 'unrotated' factors), and then a set of basic and group factors (corresponding to Cattell's 'rotated' factors).

¹ Thomson (10, p. 275) has urged much the same criticisms against the earlier analysis of similar assessments obtained from men (2). In both cases the rotated factors are expressed in the form of a 'reference vector structure' ; but we are told that the loadings should be proportional to these. On this point also see Thomson, *loc. cit.*, pp. 277f.

II. RE-ANALYSIS BY SIMPLE SUMMATION: METHODS PROPOSED

The procedure here followed differs from that adopted by Dr. Cattell in the following respects.

1. It will be remembered that Dr. Cattell's traits, like his factors, were expressed in bipolar form : but the polarity does not run in the same direction for all the traits. Thus, trait 4 is "Self-assured *versus* Submissive," while trait 15 is "Self-effacing *versus* Self-Willed." It is not surprising, therefore, to find that, in his table of coefficients, there is a *negative* correlation between these two traits amounting to -0.66 : had both traits been measured in the same direction, the correlation between them would have been $+0.66$. Dr. Cattell explains that "in the defined variable list actually given to the raters" he "deliberately reversed the polarity of items, indiscriminately" (1, p. 116). The result is that, on making the centroid computations in the usual way, Dr. Cattell's *first* factor is not a 'general factor' of the familiar positive type, but apparently bipolar : about half its saturations are negative, though the total does not add up to zero. It is well known to factorists, however, that, if the first factor is bipolar or rather quasi-bipolar, a number of difficulties, both practical and theoretical, are likely to be encountered in completing and in interpreting the factor analysis.

Accordingly, before proceeding to re-analyse the table of correlations, reversals were made in certain traits to ensure that all had a similar direction. 16 traits (marked by an asterisk in Table I below) required alteration ; as will be seen, the result is that the sum of the correlations for every trait now becomes positive. It will be noted that Dr. Cattell himself, when he comes to describe his final factors, reverses many of his traits in this fashion. All that is here proposed, therefore, is that the requisite reversals should be carried out at the beginning of the analysis instead of at the end.

2. In assessing the communalities, Dr. Cattell has followed Prof. Thurstone (8, pp. 233, 238), and taken the highest correlation in each column as indicating the self-correlation for the first factor, and the highest residual as indicating the requisite diagonal figure when calculating the later factors. This procedure, as has frequently been pointed out, tends to exaggerate both the size of the communalities and the number of the factors. As a result of his experience with this and other tables, Dr. Cattell himself remarks that he is now "convinced that the method leads to over-estimation : certainly" (he adds) "the communalities obtained in Table III" (his table of unrotated centroid factors) "err on the side of being too large" (1, p. 119).

In the following analysis the principles laid down by Burt were adopted : the communalities were kept to the minimum values required to yield the minimum number of factors, and these minimum values were reached by successive approximation. As a consequence the communalities were reduced to the figures shown in Table II below (last column, to be compared with the last column of Table III in Dr. Cattell's article) ; and the number of fully significant factors was thus reduced from 12 to 5. There can be little question that Dr. Cattell's method of assessing the communalities has introduced a number of artificial bipolar factors ; and these artificial factors in their turn have doubtless been responsible for the trouble he experienced in carrying out the rotations, since it is particularly difficult to rotate artificial factors so as to secure a clear and intelligible 'simple structure.'

3. In extracting a factor from a table of residuals by Thurstone's method, the computer has first to substitute the largest residual in each column for the diagonal value, and then to follow a somewhat elaborate and arbitrary procedure for reversing certain signs. But, apart from being unnecessarily complicated in actual application, this method has several disadvantages. Here the most obvious is that the resulting factor saturations do not add up to zero, and therefore fail to balance. Hence the classification implied by each bipolar factor must involve somewhat arbitrary lines of division.

TABLE I. OBSERVED CORRELATIONS

Trait	11	22	12	23	24	15	7	19	1	27	26	14	30	8	2
11. Suspicious ..	...	.68	.52	.59	.57	.53	.50	.43	.55	.12	.36	.51	.54	.12	.53
22. Jealous ..	.68	...	.63	.51	.72	.70	.42	.51	.49	.49	.23	.40	.43	.00	.55
*12. Spiteful ..	.52	.63	...	.63	.64	.69	.67	.72	.60	.40	.03	.24	.27	-.22	.29
23. Rigid ..	.59	.51	.63	...	.48	.44	.49	.42	.40	.05	.17	.32	.33	-.06	.19
24. Impatient ..	.57	.72	.64	.48	...	.74	.50	.59	.59	.59	.17	.38	.41	-.02	.55
*15. Self Willed ..	.53	.70	.69	.44	.74	...	.49	.71	.54	.65	-.03	.33	.30	-.31	.44
* 7. Aloof ..	.50	.42	.67	.49	.50	.49	...	.61	.65	.36	-.02	.13	.15	-.11	.24
19. Stern ..	.43	.51	.72	.42	.59	.71	.61	...	.52	.53	-.30	.06	.10	-.49	.11
* 1. Obstructive ..	.55	.49	.60	.40	.59	.54	.65	.52	...	.49	.03	.23	.25	-.01	.45
*27. Unscrupulous ..	.12	.49	.40	.05	.59	.65	.36	.53	.49	...	-.10	.18	.20	-.18	.49
*26. Anxious ..	.36	.23	.03	.17	.17	-.03	-.02	-.30	.03	-.10	...	.54	.50	.59	.45
14. Hypochondriacal ..	.51	.40	.24	.32	.38	.33	.13	.06	.23	.18	.54	...	.64	.18	.52
30. Neurotic Fatigue ..	.54	.43	.27	.33	.41	.30	.15	.10	.25	.20	.50	.64	...	.26	.49
8. Easily Upset ..	.12	.00	-.22	-.06	-.02	.31	.11	.49	-.01	-.18	.59	.18	.26	...	.39
* 2. Changeable ..	.53	.55	.29	.19	.55	.44	.24	.11	.45	.49	.45	.52	.49	.39	...
35. Immature ..	.15	.05	-.08	-.12	.09	-.13	.15	-.23	.24	.14	.28	.14	.14	.61	.46
*17. Fickle ..	.15	.11	-.02	-.23	.24	.11	.04	-.10	.41	.44	.13	.21	.18	.30	.50
6. Frivolous ..	.26	.35	.23	-.01	.51	.40	.27	.21	.57	.64	.07	.23	.22	.21	.63
*21. Awkward ..	.30	.24	.17	.09	.28	.20	.37	.20	.40	.34	.05	.05	.08	.27	.42
10. Boorish ..	.25	.25	.12	-.08	.28	.17	.22	.07	.34	.43	.08	.15	.10	.36	.49
4. Assertive ..	.22	.46	.43	.27	.50	.66	.15	.58	.10	.37	-.11	.23	.22	-.50	.12
* 9. Energetic ..	.32	.05	-.18	-.31	-.02	.11	-.39	.05	-.36	.11	-.09	.18	-.12	-.20	-.11
3. Attention-getting ..	.34	.53	.29	.15	.59	.60	.07	.32	.23	.61	.17	.38	.40	-.10	.57
28. Self-possessed ..	-.20	.07	-.05	-.17	.09	.30	-.32	.19	-.23	.26	-.20	.03	.01	-.49	-.08
33. Frank ..	-.02	.16	.00	-.10	.29	.29	-.27	.15	-.08	.27	-.04	.14	.13	-.12	.12
*18. Bold ..	-.20	.04	-.15	-.40	.14	.26	-.27	.10	-.12	.44	-.22	-.07	.00	-.21	.14
25. Eccentric ..	.09	.17	.13	.00	.23	.43	.12	.39	.16	.49	-.32	-.02	.05	-.34	.15
32. Sex ..	-.07	.20	-.02	-.16	.27	.30	-.12	.02	.05	.44	.07	.15	.11	-.04	.25
*16. Talkative ..	-.07	.19	-.15	-.25	.22	.28	-.41	-.01	-.13	.35	.13	.14	.15	.01	.31
34. Sociable ..	-.15	.11	-.27	-.37	.12	.09	-.46	-.18	-.22	.24	.14	.10	.11	.16	.27
*13. Emotional ..	.28	.38	.09	.02	.41	.34	.11	-.04	.12	.32	.53	.40	.49	.38	.58
29. Imaginative ..	-.14	-.06	-.29	-.24	-.04	-.13	-.30	-.36	-.13	.08	.20	.15	.21	.32	.24
* 5. Cheerful ..	-.50	-.25	-.46	-.58	-.16	-.15	-.57	-.30	-.38	.08	-.08	-.26	-.23	.06	-.03
*20. Indolent ..	-.20	-.10	-.30	-.47	.06	-.02	-.17	-.17	.07	.34	-.09	.00	.01	.11	.30
*31. Inartistic ..	-.06	-.12	-.18	-.10	-.07	-.11	-.10	-.08	.04	-.01	.00	-.07	-.08	.14	.04
Total ..	6.66	9.49	5.42	1.90	10.89	10.22	2.24	5.33	6.86	10.65	3.32	6.56	7.05	1.07	11.06

BETWEEN TRAITS

35 17 6 21 10	4 9 3 28 33	18 25 32	16 34 13 29	5 20 31
.15 .15 .26 .30 .25 .05 .11 .35 .24 .25 -.08 -.02 .23 .17 .12 -.12 -.23 -.01 .09 -.08 .09 .24 .51 .28 .28	.22 -.32 .34 -.20 -.02 .46 -.05 .53 .07 .16 .43 -.18 .29 -.05 .00 .27 -.31 .15 -.17 -.10 .50 -.02 .59 .09 .24	-.20 .09 -.07 .04 .17 .20 -.15 .13 -.02 -.40 .00 -.16 .14 .23 .27	-.07 -.15 .28 -.14 .19 .11 .38 -.06 -.15 -.27 .09 -.29 -.25 -.37 .02 -.24 .22 .12 .41 -.04	-.50 -.20 -.06 -.25 -.10 -.12 -.46 -.30 -.18 -.58 -.47 -.10 -.16 .06 -.07
-.13 .11 .40 .20 .17 .15 .04 .27 -.37 .22 -.23 -.10 .21 .20 .07 .24 .41 .57 .40 .34 .14 .44 .64 .34 .43	.66 .11 .60 .30 .29 .15 -.39 .07 -.32 -.27 .58 .05 .32 .19 .15 .10 -.36 .23 -.23 -.08 .37 .11 .61 .26 .27	.26 .43 .30 -.27 .12 .12 .10 .39 .02 -.12 .16 .05 .44 .49 .44	.28 .09 .34 -.13 -.41 -.46 .11 .30 -.01 -.18 -.04 .36 -.13 -.22 .12 .13 .35 .24 .32 .08	-.15 -.02 -.11 -.57 -.17 .10 -.30 .17 .08 -.38 .07 .04 .08 .34 .01
.28 .13 .07 .05 .08 .14 .21 .23 .05 .15 .14 .18 .22 .08 .10 .61 .30 .21 .27 .36 .46 .50 .63 .42 .49	-.11 -.09 .17 -.20 -.04 .23 -.18 .38 .03 .14 .22 -.12 .40 .01 .13 -.50 -.20 .10 .49 .12 .12 .11 .57 .08 .12	-.22 -.32 .07 -.07 -.02 .15 .00 .05 .11 -.21 .34 .04 .14 .15 .25	.13 .14 .53 .20 .14 .10 .40 .15 .15 .11 .49 .21 .01 .16 .38 .32 .31 .27 .58 .24	-.08 -.09 .00 -.26 .00 .07 -.23 .01 .08 .06 .11 .14 -.03 .30 .04
... .49 .46 .44 .70 .4968 .35 .54 .46 .6844 .58 .44 .35 .4455 .70 .54 .58 .55 ...	-.51 -.33 .01 .50 .13 -.21 .19 .23 .10 .08 .05 .06 .45 .02 .14 -.10 .16 .17 .29 .15 -.18 .18 .21 .25 .06	-.18 .23 .06 .25 .10 .26 .30 .28 .35 .06 .20 .09 .06 .02 .25	-.04 .09 .16 .08 .22 .33 .26 .28 .27 .28 .34 .20 .14 .04 .27 .04 .16 .23 .22 .06	-.02 .22 .10 .17 .63 .20 .12 .51 .06 -.07 .28 .52 .05 .27 .17
-.51 .21 .05 .10 .18 -.33 .19 .06 .16 .18 -.01 .23 .45 .17 .21 -.50 .10 .02 .29 .25 -.13 .08 .14 .15 .06	43 .63 .64 .52 .4330 .65 .53 .63 .3047 .50 .64 .65 .4755 .52 .53 .50 .55 ...	.42 .39 .27 .69 .28 .33 .51 .45 .49 .64 .38 .50 .61 .36 .32	.44 .25 .31 .11 .62 .54 .33 .16 .58 .47 .56 .18 .61 .49 .22 .11 .74 .61 .46 .11	.10 .07 .10 .68 .11 .00 .22 .27 .02 .42 .19 .02 .44 .25 .14
-.18 .25 .30 .06 .06 .23 .10 .28 .20 .02 .06 .26 .35 .09 .25	.42 .69 .51 .64 .61 .39 .28 .45 .38 .36 .27 .33 .49 .50 .32	48 .45 .4821 .45 .21 ...	.75 .67 .42 .26 .32 .11 .13 .23 .42 .51 .36 .16	.68 .51 .10 .13 .31 .01 .33 .27 .12
.04 .22 .27 .14 .16 .09 .33 .27 .04 .23 .16 .26 .34 .27 .22 .08 .28 .20 .04 .06	.44 .62 .58 .61 .74 .25 .54 .47 .49 .61 .31 .33 .56 .22 .46 -.11 .16 .18 .11 .11	.75 .32 .42 .67 .11 .51 .42 .13 .36 .26 .23 .16	78 .60 .23 .7851 .29 .60 .5134 .23 .29 .34 ...	.63 .36 .20 .67 .42 .10 .27 .27 .11 .31 .42 .03
-.02 .17 .12 .07 .05 .22 .63 .51 .28 .27 .10 .20 .06 .52 .17	.10 .68 .22 .42 .44 -.07 .11 .27 .19 .25 -.10 .00 .02 .02 .14	.68 .13 .33 .51 .31 .27 .10 .01 .12	.63 .67 .27 .31 .36 .42 .27 .42 .20 .10 .11 .03	40 .11 .4027 .11 .27 ...
2.74 7.04 10.21 5.58 6.59	6.87 2.67 11.85 3.94 7.31	7.16 5.85 6.78	8.79 7.07 10.34 2.63	1.83 5.26 1.12

Burt's method is based on the principle of weighting certain rows and columns, so that the factor-variance is maximized. This is most readily achieved by first rearranging the residuals so that the table containing them is divided into positive and negative quadrants corresponding with the best-fitting bipolar factor.¹ In passing it may be observed that the device of rearrangement greatly facilitates an understanding of the relations between the variables in question and the choice of an appropriate procedure : it is, I venture to suggest, too often neglected.

4. Cattell attempts no interpretation of the bipolar factors obtained by his centroid analysis. They are used merely as a convenient starting-point for a series of 'blind' rotations (1, p. 119). A graphical method was employed ('single plane' procedure, 9, p. 216) ; and the rotations were continued, on a trial-and-error basis, until a satisfactory set of group factors was achieved.

The method adopted in the present analysis falls into two stages. First, a preliminary set of non-overlapping group-factors is obtained. For these the essential lines of demarcation between the groups are based on the points of demarcation between the positive and negative sections of the bipolar factors. Secondly, a matrix of overlapping group factors is obtained by rotation, the requisite rotation matrix being obtained, not by a rough graphical procedure, but by direct arithmetical calculation, based on the preliminary group factor matrix. With a suitably planned enquiry this may complete the analysis. But in other cases a supplementary process of successive approximation is necessary. Here, owing to certain peculiarities in the choice of traits (e.g., the deliberate duplication of a good many of them), a considerable amount of supplementary work was required.

5. Of the 36 traits appearing in Dr. Cattell's correlation table all except one are traits of 'personality' in the narrower sense of that term : i.e., they refer to 'affective' or 'conative' characteristics. The exception is the last trait of all (no. 36). This is described as an intelligence test, and is not included in Cattell's main list. It is the only ² 'cognitive' trait in the whole battery. It is therefore likely to include none of the general or group factors shared by the other traits, and is almost bound to involve a large specific factor, quite unrelated to the rest. If retained in the analysis, the presence of a trait so different from the remainder will only upset the determination of the common factors characteristic of the remaining traits. This seems confirmed by the results of Cattell's own figures. In his table of observed correlations the coefficients for this trait are nearly all non-significant : only one has a numerical value higher than 0.20. In the rotated factor-matrix this trait yields a significant saturation (0.60) for only one of the factors (the factor termed 'intelligence'), and for this factor only one other trait has a loading of over 0.25 (no. 17, "persevering versus fickle"). Now factor-analysis is essentially a search for common factors, i.e., general factors or group factors shared by more than one trait. If a trait belongs to an entirely different class from the rest, and has no factor in common with them, it is plainly absurd to retain it in a factor-analysis. Trait 36 has therefore been dropped.

Table I shows the correlations between the 35 traits, modified as explained above, and rearranged in the order required for the final analysis. With the reversals proposed it becomes much easier to rearrange the traits according to a more intelligible scheme ; and in the present case this serves to indicate, even before the computations have been completed, what are the main structural relations and groupings of the traits selected.

¹ The method of 'simple summation' was introduced as an approximate device for ascertaining the factors required by Pearson's 'least squares' procedure, which entailed a method of 'weighted summation.' In place of the *differential* weights the simplified procedure substitutes *equal* weights, i.e., weights that are numerically the same and differ only in sign (usually ± 1). It seems probable that, with the present correlation table, not much accuracy is lost by this substitution. There is, however, a further point in which the working procedure proposed differs from that proposed by Pearson : minimal self-correlations are substituted for unity in the leading diagonal. Strictly speaking, as Burt has pointed out, this change should lead to the substitution of $|R_c R^{-1} - \lambda I|$ for $|R^{-1} - \lambda I|$ in Pearson's determinantal equation (R_c denoting the reduced correlation matrix : for proof cf. 5). It follows that when this substitution is not carried out and the simplified procedure adopted, the chi-squared test of significance becomes only approximate.

² As will be seen from the correlations and saturations, no. 10 ("boorish versus cultured") does not appear to be a cognitive trait as might at first be thought.

(a) *Bipolar Analysis.* A fresh analysis by simple summation, along the lines described above, was first carried out. With the procedure proposed five factors at most seem sufficient to give a plausible fit to the observed correlations. The saturations obtained are shown in Table II. Judged by the chi-squared test¹ the residuals then remaining have little or no statistical significance. Indeed, if we ignore two large residuals (due apparently to exceptional features in three of the trait ratings, those for nos. 7, 21, and 31), those remaining after only *four* factors have been extracted would seem to be devoid of significance: their standard deviation is ± 0.059 , and the standard deviation of a zero correlation with only 240 persons is ± 0.065 .

Taken just as they stand, the four unrotated factors thus obtained yield an intelligible and suggestive classification of the 35 traits; and the classification so reached has the further advantage that it has been procured quite mechanically without the arbitrary element involved in successive rotations.

(b) *Group Factor Analysis.* The classification indicated by the bipolar factors forms the basis for the group factor analysis. On examining Table II, it will be observed that the saturations in both sections (positive and negative) of the first bipolar factor, and in the first two subsections of the second bipolar factor, are, on the whole, fairly large: the remaining saturations, with two exceptions only, are below 0.60. This suggests that the group factor analysis may well begin with three large groups, namely, traits 11 to 27, 26 to 10, and 4 to 31 as arranged in the table. Any smaller group factors can be added or substituted later, if the residuals suggest such changes.

An analysis into one 'basic' and three non-overlapping 'group' factors was therefore first carried out along these lines (for details of the working procedure, see 5). Then, to locate the points of overlapping, a rotation matrix was calculated in the usual way²; and the bipolar factors rotated accordingly.

The residuals suggested the need for 8 smaller group factors, corresponding to the 8 subsections of the third bipolar factor. On making these additions, it then appeared that one of the larger group factors (that for traits 26 to 10) was no longer necessary. No doubt some such factor exists; but, after removing the remaining factors, the residuals on which it would have been based proved to be, with this small sample, too low to be significant.

The values thus obtained for the factor-saturations were checked and adjusted by successive approximation. It was found that, with slight allowances for overlap and with a few small supplementary factors to account for the cases of closely identical traits, a fairly good fit could be secured. The complete set of saturations finally reached is shown in Table III.

The advantages of the procedure here adopted will be obvious. The initial group factors are calculated directly from the original table of correlations; and the lines of division between the groups are determined, not by the somewhat arbitrary selection of the computer, but quite automatically by the sign-pattern of the preliminary bipolar matrix. The rotation is thus based directly on the relations between

¹ With the simplified procedure here adopted, this and other tests of significance have only an approximate validity: cf. footnote 1, p. 208.

² The mode of calculation is as follows. With three group factors and one basic factor, both bipolar and group factor matrices are first reduced by summation to square (4×4) matrices, F_1 and G_1 say. From G_1 a fresh covariance matrix, $R_1 = G_1 G_1'$, is constructed, and factorized by simple summation, yielding say F_2 . It follows that, if we write $G_1 T = F_2$, T must be exactly orthogonal. We accordingly calculate $T = G_1^{-1} F_2$, and $G_2 = F_1 T'$. G_2 will then be the required matrix of rotated factors, and will consist of overlapping, orthogonal group factors, fitting the observed correlation matrix, R_1 , with the same degree of accuracy as F_1 .

TABLE II. GENERAL AND BIPOLAR FACTORS

Traits				Factors					Sum of Squares							
				I	II	III	IV	V								
IIA. DYSTHYMIC																
IIIA. Organized (Schizothymic)																
IVA. Paranoid Group																
11.	Suspicious	..	..	·47	—·60	—·12	—·28	—·18	·70							
22.	Jealous	..	..	·66	—·37	—·29	—·25	+·13	·74							
12.	Spiteful	..	..	·40	—·56	—·55	—·06	+·09	·78							
23.	Rigid	..	..	·16	—·57	—·41	—·36	—·11	·66							
24.	Impatient	..	..	·74	—·31	—·32	—·12	—·03	·76							
IVB. Compulsive Group																
15.	Self Willed	..	..	·71	—·20	—·55	+·07	+·06	·84							
7.	Aloof	..	..	·19	—·65	—·43	+·15	+·09	·67							
19.	Stern	..	..	·40	—·28	—·74	+·23	—·08	·85							
1.	Obstructive	..	..	·49	—·60	—·15	+·29	—·15	·72							
27.	Unscrupulous	..	..	·73	—·04	—·19	+·43	+·09	·76							
IIIB. Unorganized (Morbid Cyclothymic)																
IVC. Anxiety Group																
26.	Anxious	..	..	·26	—·24	+·46	—·63	+·09	·75							
14.	Hypochondriacal	..	..	·45	—·25	+·09	—·42	—·11	·46							
30.	Neurotic Fatigue	..	..	·49	—·24	+·09	—·46	—·17	·55							
8.	Easily Upset	..	..	·11	—·23	+·76	—·21	+·06	·68							
2.	Changeable	..	..	·76	—·30	+·31	—·09	+·03	·78							
IVD. Adolescent Hysteria Group																
35.	Immature	..	..	·23	—·38	+·67	+·25	+·31	·80							
17.	Fickle	..	..	·49	—·09	+·48	+·40	—·07	·64							
6.	Frivolous	..	..	·70	—·18	+·23	+·42	+·04	·75							
21.	Awkward	..	..	·39	—·20	+·28	+·25	—·33	·44							
10.	Boorish	..	..	·47	—·29	+·38	+·39	+·24	·65							
IIB. EUTHYMIC																
IIIC. Organized (Volitional)																
IVE. Dominating Group																
4.	Assertive	..	..	·49	+·27	—·67	—·24	—·02	·82							
9.	Energetic	..	..	·22	+·75	—·17	—·07	+·17	·67							
3.	Attention-getting	..	..	·81	+·20	—·16	—·11	+·02	·74							
28.	Self-possessed	..	..	·30	+·72	—·41	—·08	+·03	·79							
33.	Frank	..	..	·50	+·54	—·09	—·08	—·11	·57							
IVF. Impulsive Group																
18.	Bold	..	..	·51	+·73	—·06	+·20	—·02	·83							
25.	Eccentric	..	..	·40	+·25	—·31	+·26	—·24	·45							
32.	Sex	..	..	·46	+·28	—·03	+·02	+·26	·36							
IIID. Unorganized (Healthy Cyclothymic)																
IVG. Cordial Group																
16.	Talkative	..	..	·61	+·64	+·10	—·11	+·04	·80							
34.	Sociable	..	..	·50	+·60	+·27	—·07	+·18	·71							
13.	Emotional	..	..	·71	+·17	+·27	—·35	—·03	·72							
29.	Imaginative	..	..	·18	+·26	+·36	—·08	—·08	·25							
IVH. Facile Group																
5.	Cheerful	..	..	·17	+·77	+·24	+·11	+·21	·73							
20.	Indolent	..	..	·38	+·34	+·40	+·44	—·19	·65							
31.	Inartistic	..	..	·08	+·06	+·26	+·16	—·22	·15							
Sum of Squares				..	..	..	..	..	..	8·37	6·48	4·97	2·63	·77	23·22	
Contribution to Variance				..	..	..	..	..	..	..	23·9%	18·5%	14·2%	7·5%	2·2%	66·3%

the bipolar and group factor matrices. The conscious or unconscious element of choice that intervenes when a 'primary structure' of 'oblique' or correlated factors is attained by successive rotations, without any objective guidance, can thus be eliminated. Finally, it may be noted that the entire calculations required the equivalent of a month's full-time work or rather less—scarcely a twentieth of the time required with Cattell's procedure.¹

III. RESULTS OBTAINED: INTERPRETATION OF FACTORS

Correspondence of Results. The results obtained by the procedures here suggested, i.e., the general-and-bipolar factor matrix shown in Table II and the basic-and-group factor matrix shown in Table III, have now to be compared with the corresponding results obtained by Cattell's procedure, i.e., with the set of 'unrotated factors' given in his Table III (1, p. 120) and the final set of 'rotated factors' given in his Table IV (*ibid.*, p. 122). There is, naturally enough, a good deal of similarity; but the changes are still more striking. Accordingly, so far as the figures differ, we have now to enquire whether the concrete psychological factors indicated by the new results are more plausible or less plausible than Cattell's own factors as he interprets them.

Comparison of Unrotated Factors If we reverse the signs in the appropriate rows of Cattell's 'unrotated factor matrix' (his Table III), so that all the saturations in his first column become positive, with consequent changes in the other columns, then, apart from a few notable exceptions, the saturations in the first four columns of that table do not diverge very greatly from the saturations in the first four columns of the corresponding table here (Table II, p. 210). Cattell's factor II, however, becomes our factor III; and his factor III our factor II. The most important of the detailed changes occur in the later columns; but even here the same general pattern is discernible. The bipolar factors therefore are not greatly altered in their general nature.

In both his table and mine the first three factors exhibit a rough analogy with the 'general' and 'specific' factors obtained by Burt in his analysis of affective and conative traits (4, 6). Thus, the first factor is roughly analogous to Burt's factor of 'general emotionality'; the second to his 'euphoric' and 'dysphoric' factor of 'pleasurable *versus* unpleasurable emotions'; and the third to his 'sthenic' and 'asthenic' factor (objective or extravertive *versus* subjective or introvertive emotions). But there is this important difference. Burt's earlier investigations, particularly those carried out with children, were concerned to determine (so far as possible) the factors underlying the *innate* orectic disposition—the more important 'temperamental' factors: they were therefore based as a rule on assessments of the 'primary emotions.' Cattell's traits are concerned rather to describe the young adult as he now is, and include the results of environment, habit-formation, and conflict or harmonious growth. This, as we shall see, naturally tends to modify the amount of correspondence.

1. *General Factor.* The first factor obtained by simple summation has positive saturations for every trait. Evidently, therefore, the purpose of reversing the traits has been successfully achieved: except perhaps for two doubtful cases (7 and 28) every trait is now measured in the same direction. If we allow for the consequent change of signs, the figures

¹ An even closer fit could be obtained by continuing the process of successive approximation still further, and then making a fresh rotation based on a larger number of group factors. But there is no ground for seeking an extremely close fit, when it is clear that the supplementary factors indicated *a priori* by the nature of certain pairs of traits are nevertheless bound to be non-significant with so small a sample.

TABLE III. BASIC AND GROUP FACTORS

Traits	Factors										
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
IIA. DYSTHYMIC											
IIIA. Organized (Schizothymic)											
IVA. Paranoid Group											
11. Suspicious ..	·47	·30	...	·54	...	...	...	—·39	...	...	...
22. Jealous ..	·64	·45	...	·24	...	...	...	...	...	...	...
12. Spiteful ..	·27	·72	...	·29	...	...	...	...	...	...	...
23. Rigid ..	·09	·77	...	·51	...	...	...	·34	...	...	...
24. Impatient ..	·67	·52	...	·05	...	...	...	...	...	...	...
IVB. Compulsive Group											
15. Self-willed ..	·66	·53	...	·26	·08	...	...	...	...	...	...
7. Aloof ..	—·11	·84	...	·14	·16	...	...	...	...	...	...
19. Stern ..	·29	·73	...	·19	·20	—·32	...	...	...	...	...
1. Obstructive ..	·40	·63	—·32	...	·32	...	·31	...	...	...	...
27. Unscrupulous ..	·89	·12	...	—·36	·48	—·41	...	...	...	...	...
IIIB. Unorganized (Morbid Cyclothymic)											
IVC. Anxiety Group											
26. Anxious ..	·28	—·27	...	·31	...	·79	...	...	...	...	...
14. Hypochondriacal ..	·46	...	—·19	·33	...	·43	...	...	...	...	...
30. Neurotic Fatigue ..	·47	...	...	·30	...	·50	...	...	...	...	...
8. Easily Upset ..	·16	—·32	...	...	...	·53	·28	—·69	...	...	...
2. Changeable ..	·72	·37	...	...	...	·38	·28	...	...	...	...
IVD. Adolescent Hysteria Group											
35. Immature ..	·26	—·22	—·36	...	...	·33	·72	...	...	...	...
17. Fickle ..	·35	—·15	...	...	...	...	·64	—·23	·32	·27	·38
6. Frivolous ..	·64	·11	...	...	...	...	·52	...	...	...	...
21. Awkward ..	·38	...	...	...	...	...	·44	...	...	...	...
10. Boorish ..	·40	...	...	...	...	...	·74	...	...	...	...
IIB. EUTHYMIC											
IIIC. Organized (Volitional)											
IVE. Dominating Group											
4. Assertive ..	·36	·36	·40	·47	...	...	—·32	·35	...	...	...
9. Energetic ..	·15	...	·49	—·34	...	—·33	...	·15	...	...	...
3. Attention-getting ..	·73	...	·37	...	...	...	...	·14	...	...	...
28. Self-possessed ..	·22	...	·58	...	...	—·28	...	·26	...	...	...
33. Frank ..	·30	...	·65	...	...	...	...	·35	...	...	...
IVF. Impulsive Group											
18. Bold ..	·28	...	·74	—·62	...	—·17	...	...	·31	...	...
25. Eccentric ..	·31	...	·29	...	...	—·46	...	...	·34	...	...
32. Sex ..	·27	...	·49	...	...	...	...	...	·03	...	...
IIID. Unorganized (Healthy Cyclothymic)											
IVG. Cordial Group											
16. Talkative ..	·33	—·31	·68	...	...	...	...	...	...	·37	·36
34. Sociable ..	·26	—·32	·57	...	...	...	...	...	...	·39	...
13. Emotional ..	·52	...	·39	...	...	·43	...	...	...	·46	...
29. Imaginative ..	·19	—·30	·18	...	...	...	...	...	...	·24	...
IVH. Facile Group											
5. Cheerful ..	·10	—·49	·57	—·41	—·38	...	...	...	...	·30	·04
20. Indolent ..	·17	...	·52	—·38	...	...	·37	—·42	...	·33	·44
31. Inartistic ..	·06	...	·07	—·24	...	...	...	...	...	...	·27

Note.—The small supplementary factors of narrow range needed to improve the fit (see p. 211) have been omitted from the above table.

are closely similar to those given by Cattell (*loc. cit.*, column 1) : the small discrepancies are due to the modification of the communalities as a result of successive approximation.¹ The factor accounts for 25 per cent. of the total variance.

Since most of the traits describe 'emotional' characteristics (in McDougall's sense of that term), it may seem tempting, as already suggested, to identify this factor with 'general emotionality.' The latter is no doubt largely responsible for the positive correlations. But the phrase 'general emotionality,' like the phrase 'general intelligence,' was intended to describe an estimate of an innate characteristic. And here, as we have seen, the traits as actually assessed must represent the joint products of both innate and acquired characteristics—of extra-personal and intra-personal influences operating on the individual's inborn disposition rather than that inborn disposition by itself. The factor might therefore perhaps be more accurately described as 'emotional instability,' using that term in its ordinary sense to designate a lack of balanced personality, arising from emotional causes. This is confirmed by a glance at the larger saturations themselves : 'emotional,' 'impatient,' 'frivolous,' 'changeable,' 'self-willed,' 'attention seeking,' 'unscrupulous,' all have saturations over 0.70.

2. First Bipolar Factor. The second factor accounts for 18 per cent. of the variance. It divides the 35 traits into two unequal groups—20 with negative and 15 with positive signs. The highest saturation of all (-0.76) is assigned to 'Cheerfulness'; and, with scarcely a single exception, all the traits characterized by pleasurable feeling-tone belong to the second (positive) section and all the traits characterized by unpleasurable feeling-tone belong to the first (negative) section.

Nevertheless, the simple contrast between pleasurable and the reverse does not express the most conspicuous difference between the two groups. The traits with the largest saturations appear to be concerned, not so much with pleasant or unpleasant feeling-tone as such, but rather with ease or difficulty of emotional adjustment. The term that best covers the 'cheerful,' 'bold,' 'energetic,' 'sociable,' 'self-possessed,' 'talkative' person is perhaps 'euthymic'; similarly the person who is 'aloof,' 'suspicious,' 'obstructive,' 'rigid,' and somewhat 'immature,' 'changeable,' 'neurotic,' and 'anxious,' may be called 'dysthymic.'

It is unfortunate that the number of traits in the two sections is so unequal. The inequality necessarily tends to increase the size of the saturations in the smaller group and to decrease the size of those in the larger group,² since otherwise the total would not be zero. As we shall see in a moment, it also seems to distort the results of the group factor analysis.

3. Second Bipolar Factor. Factor III accounts for 14 per cent. of the variance. It divides each section of Factor II into two contrasted subsections. The traits that exhibit the highest positive saturations are 'stern,' 'assertive,' 'self-willed,' 'spiteful,' 'aloof,' and 'self-possessed': this suggests an 'organized' or 'hardened' personality, i.e., one capable of deliberate and volitional insistence, possessing an objective or realistic standpoint, and a strong or sthenic attitude. The traits that exhibit the highest negative saturations are 'being easily upset,' 'immature,' 'anxious,' 'indolent,' and 'imaginative': this suggests an undeveloped or ill-organized type of person who is at the mercy of his feeling and fancies, one who possesses a subjective standpoint and a weak, asthenic attitude. Accordingly the phrase 'organization *versus* lack of organization' will perhaps serve to designate this factor.³

¹ This was carried out with the full correlation table. With large tables much time may be saved if Burt's device of factorizing *summed correlations* is used (5). Here, however, this would have precluded a study of the residuals and any changes in the lines of partition.

² The effects of the lack of balance are still more obvious in Cattell's table. The result is that several of the traits in the larger group—'stern,' 'self-willed,' and 'unscrupulous'—are separated from those to which they naturally belong, such as 'aloof,' 'obstructive,' and 'spiteful,' and are thrown into the other group. On the other hand, 'inartistic,' which with my analysis has a very small saturation in the 'euthymic' section (-0.06), is in Cattell's analysis rightly removed to the other section ($+0.15$).

³ The term 'organized' does not necessarily imply a well-balanced personality, or one organized in harmonious adjustment with external requirements. A personality can be 'organized' for ill as well as for good. As here employed, the word would cover such extreme manifestations as those described by psychiatrists as 'organized' delusions or impulses, or by lawyers as 'system.'

4. *Third Bipolar Factor.* Factor IV accounts for 7 per cent. of the variance. It subdivides each of the four sections of factor III ; and thus yields eight small groups, each containing 3 to 6 traits. As we shall find in a moment, these all reappear in the group factor analysis ; and their interpretation can accordingly be best discussed in connection with the rotated factors obtained by Cattell.

Comparison of Rotated Factors. One or two comments are first required on the general nature of the group factors here obtained as contrasted with those secured by Cattell for his final interpretation. First, it is to be noted that, unlike the factors ordinarily secured by rotation, Cattell's group factors are all bipolar. On the other hand, those resulting from the present analysis have (except for a few intelligible cases of overlap) positive saturations throughout.¹

Naturally, if it is defined by the factor-measurements, every factor can be regarded as bipolar, since all such measurements are of necessity both positive and negative : thus, just as Cattell expresses his personality traits and his personality factors in bipolar form—'Hypochondriacal *versus* Not so,' 'Marked interest in Sex *versus* Slight Interest in Sex,' and the like—so we could, if we did not mind the cumbersome descriptions that would result, describe even cognitive abilities and cognitive factors in the same antithetical fashion : we could regularly re-label Intelligence as 'Intelligence *versus* Lack of it' (as Cattell does in discussing his factor IX), and any other ability or propensity as 'So-and-so *versus* Not-So-and-so.' But this only illustrates the artificiality of the device, and serves to show how easily such a procedure may lead to erroneous inferences. To begin with, it entirely obscures the distinctions between general and non-general factors, and between group factors and genuine bipolar factors (i.e., factors with negative factor-saturations, not merely negative factor-measurements). Further, when this antithetical principle is enforced for every trait or factor there is a strong tendency to confuse contradictories and contraries.² Quite apart from this confusion, however, many of the alleged opposites are not true opposites at all : for example, one would expect the opposite of 'infantile sthenic emotionality' to be 'adult asthenic emotionality' ; but it is not : it is described as 'phlegmatic frustration tolerance' : and so of many of the factors. Moreover, the bipolarity begs many questions which should be settled empirically : for instance, it is assumed that, since 'grief' is at the opposite pole to 'cheerfulness,' no separate rating is needed for 'grief' as distinct from 'cheerfulness.' Consequently, the possibility of a *positive* correlation between the tendencies to grief and to cheerfulness is excluded from the outset ; yet, as we know, the empirical evidence shows that highly emotional persons are commonly prone to both. Finally, if, in spite of these objections, the intention was to have all factors bipolar, the appropriate procedure should surely have been a summational factor analysis so arranged as to permit limited as well as general bipolar factors to emerge.

There is a further defect about Cattell's proposed scheme of factors. The bipolar factors which he eventually reaches exhibit no kind of co-ordination or systematic relations. If, as has been urged, factor analysis is merely a quantitative method of classification, one would expect that, given suitable traits, the results would show some tendency towards a systematic hierarchy³ of classes—a *summum*

¹ Most of the negative saturations could be suppressed by introducing further positive factors. The majority are due to the omission of the group factor for traits 26 to 10 ; a few are due to the way certain traits have been selected.

² In logic pairs of terms like 'white' and 'not-white,' 'moral' and 'non-moral,' are called 'contradictories,' ; pairs like 'white' and 'black,' 'moral' and 'immoral,' are called contraries. Confusing the two is a well-known fallacy. In psychology, antitheses such as 'intelligent' and 'not intelligent,' 'marked sex interest' and 'slight sex interest,' do not imply qualitative contrasts, as do (for example) 'pleasant' and 'painful,' 'joy' and 'sorrow,' 'obstructive' and 'co-operative.'

³ Cattell has himself (3, p. 280) commented on the 'hierarchical' arrangement of factors in Burt's schemes, or (as he has termed it) the 'genealogical' pattern, where a broader group (or bipolar) factor becomes, as it were, the parent of two or more such factors, which in turn may each bring forth two or more still smaller factors. He apparently would doubt whether personality factors can really be arranged in this way. Whether one accepts the principle or not, must of course depend on whether one accepts the notion of some kind of hierarchical organization in the mind. On this see McDougall's chapter on the 'Structure of Character' in his *Social Psychology*, 1936 ed., pp. 431f.

genus, which included subordinate genera, narrower species within these subordinate genera, and sub-species within the species. Instead we are confronted (as in the results so often presented by empirical factorists) with a mere miscellaneous assortment. No doubt there are dangers in forcing on the results a hasty systematization at all costs which is not naturally suggested by the data. Nevertheless, a method which almost automatically entails a complete absence of structural relations is surely at fault.

As a result of these two peculiarities in Cattell's working assumptions, it follows that Cattell's list can have no general factor, representing the *summum genus* of all the traits, and possessing positive saturations throughout. The consequence is that his group factors have to do the work of the general factor; and are thus forced to overlap with each other enormously.

His factor 1, for example, instead of being confined to a small group of traits, as we should expect in a 'simple structure,' has as many as 18 saturations over ± 0.10 ; his factor 2 has 20. Each of them has thus more overlapping saturations than non-overlapping. And much the same is true of the other factors. Many of the traits have saturations over ± 0.10 for more than half the factors. With eleven widely overlapping factors of this kind the effect of a general factor is readily accounted for without actually introducing a general factor itself; but as a result all the factors lose much of their distinctness. Take, for example, factor 4 ('dominance'): this, like my own factor of dominance, has comparatively high saturations for such traits as 'energy,' 'assertion,' 'attention-getting,' 'self-possessed,' 'frank,' 'tough, unshakable poise'; but it also has saturations over 0.30 for 'neurotic fatigue,' 'hypochondria,' and 'depression.' Cattell explains this on the ground that the dominant woman is apt to get her way by behaving in a neurotic and hypochondriacal manner. But his observed correlations hardly bear this explanation out: the cross-correlations between the two sets of traits average only about 0.16. The results of this wide overlapping are shown both in the correlations between the factors and in the similarity of names or descriptions assigned to them. Ten of the correlations are over ± 0.35 . Three of the factors are described as expressing 'immature' or 'infantile' characteristics; two as expressing 'cyclothymic' characteristics.

On the other hand, with the procedure adopted here, the relations between the summational factors and the group factors ensure that (except for the small supplementary factor-saturations due to the duplication of certain traits or to narrow overlaps) the group factors finally reached are virtually uncorrelated.

Interpretation of Group Factors. Since the group factors here obtained are directly related to the factors (or sections of factors) obtained in the preliminary bipolar analysis, there is little to add in regard to their concrete interpretation.

1. *Basic Factor.* The 'basic' factor, shown in Table III, is evidently much the same as the 'general' factor, shown in Table II. Although the absolute size of the saturations is (as usual) somewhat reduced, their relative size remains for the most part the same as before. One trait—'aloof'—has a small negative saturation; but that is quite in keeping with the interpretation proposed. As we have already observed, Cattell's mode of rotation excludes the discovery of any general factor of this kind, though if personality as a whole is to be consistent we should certainly expect at least some small common factor running through all the traits.

2. *Broad Factors.* We might perhaps have expected two fairly broad factors corresponding to the positive and the negative halves of the first bipolar factor. Actually the residuals here only suffice to establish one broad factor corresponding to the 'euthymic' section of the first bipolar and a somewhat narrower factor corresponding to the 'schizothymic' or introvertive section of the second bipolar. This is due to the fact that, as we have already noted, the number of traits in the two sections of the first bipolar is unfortunately so unequal. Such distortions of the group factor matrix are familiar to every factorist who has attempted to factorize a battery of tests or traits in which the several sub-classes are disproportionately represented.

The broader of these two group factors ('euthymic') seems to correspond with Cattell's factor 5 (which he names 'surgency'); the narrower ('schizothymic') corresponds with his factor 2 ('cyclothymia *versus* schizothymia')—the reversal of the description is due to the reversal of the relevant traits).

3. *Narrow Factors.* We come finally to the eight group factors based on the eight subsections of the third bipolar factor. For some of these factors the residual correlations are not very large, and the evidence therefore not highly significant. Yet I fancy those who know women students in this country (as well as in America) will agree that every one of the types discovered by the analysis is familiar from everyday observation.

The first two types (i and ii) may be fairly regarded as subdivisions of the negativistic or so-called 'schizophrenic' type. This, particularly among children and young people, is apt to take two main forms, leading to a 'suspicious' subtype and a 'compulsive' subtype respectively. The former is characterized largely by obsessive emotions; the latter by obsessive actions.

(i) *Paranoid Type.* The person who is 'suspicious,' 'jealous,' 'spiteful,' 'unadaptable, or rigid,' corresponds with what is commonly described in psychiatric literature as paranoid. Among Cattell's rotated factors that numbered 11 has significant factors for all the traits in this group (and one or two others).¹

(ii) *Compulsive Type.* Judged by the saturations, the second group is 'self-willed,' 'stern,' 'obstructive,' 'unscrupulous,' and 'aloof.' These are usually regarded as the dispositional characteristics of those in whom obsessions and compulsive actions readily develop. So far as the numerical results are concerned, this factor seems roughly identifiable with Cattell's rotated factor 3: he describes it as 'immature dependent character *versus* positive character integration.'² The description, however, does not seem altogether appropriate. It is difficult to see anything 'immature' or 'dependent' in being 'obstructive' or 'unscrupulous.' Moreover, trait 35 is designated 'dependent, immature'; and trait 24 also describes the 'emotionally mature' and their opposite (the 'immature' presumably): yet Cattell's factor 3 has a correlation of practically zero with these two traits.

(iii) *Anxiety Type.* This group is characterized more especially by 'anxiety,' 'neurotic fatigue,' 'hypochondria.' The factor seems identical with Cattell's rotated factor 1. He describes it as 'neurotic general emotionality *versus* emotionally mature stable character.' Since an 'integrated character' must surely be a 'stable character,' and since 'mature' is the opposite of 'immature,' Cattell's wording would suggest that this factor and the last (factor 3) are closely allied. Actually, however, his saturations show little or no overlapping: and, curiously enough, his factor 1 has very small saturations for traits 35 ('immature'), and 24 ('emotionally mature'). On the other hand, his largest saturations are for trait 26 ('worrying, anxious') and 30 ('neurotically fatigued').

(iv) *Adolescent Hysteria.* The type of person indicated by this factor is apparently distinguished by two main features: he is 'immature,' 'awkward,' and 'boorish'; and he is 'frivolous' and 'fickle.' These are the emotional characteristics commonly singled out as characteristic of the undeveloped, adolescent personality, prone to the kind of emotional instability that is often dubbed 'hysterical.' Cattell's factor 6 has large saturations for most of these traits. As it also has moderately large saturations for trait 31 ('lack of artistic feeling') and trait 29 ('practical, logical, *versus* sensitively imaginative'), he describes it as a factor for 'boorishness *versus* trained socialized cultured mind,' or (on another page)

¹ It is a prevalent custom among psychiatrists to borrow labels, introduced originally to describe abnormal or pathological conditions, and to employ them for the purpose of designating perfectly normal types observable among perfectly normal persons. Psychologists, as a rule, strongly deprecate this practice. Many of Cattell's traits, however, relate specifically to pathological tendencies or symptoms (e.g., 'hypochondriacal,' 'neurotic fatigue,' 'immature,' etc.); and it would seem that, in his approach to the study of personality factors, he started by taking the 'clinically distinguishable forms' and the 'principal pathological syndromes' as offering the clearest points of departure (cf. 3, chaps. ii and iii). Since my object here is to link up the types revealed by the factorial analysis with the types already recognized and named, it would seem best to accept the labels most commonly used.

² To conform with the direction adopted for the various traits in this study, I have ventured to reverse the order in which the antithetical tendencies are specified in Cattell's description of his factor 3.

'general education' *versus* the lack of it. Once again his descriptions do not seem wholly suitable: because a man is 'practical or logical' rather than 'imaginative,' must we say that he lacks a 'trained cultured mind'? And should we not expect a factor for 'general education' and a 'trained cultured mind' to have a correlation with intelligence of rather more than 0.02?

(v) *Dominating Type*. The nature of the fifth group factor, which has high saturations for such traits as 'energy,' 'assertion,' 'attention-getting,' 'frankness,' and 'tough, unshakable poise,' is not difficult to determine. Such characteristics are common in the popular conception of the man of 'strong will'—the masterful, resolute, dominating type. The factor involved corresponds fairly well with Cattell's factor 4, which he appropriately terms 'dominance.'

(vi) *Impulsive Type*. The traits forming this factor are designated 'adventurous or bold' (as contrasted with 'cautious or retiring'), 'eccentric or unconventional,' and 'marked interest in the opposite sex.' This combination suggests a sthenic personality, somewhat like the last, but one governed not so much by deliberate will or well-established purpose as by daring impulse and the capricious whim of the moment. Cattell's factor 7 is named 'adventurous cyclothymia'; but its saturation for trait 18 ('adventurous, bold') is only 0.01, and for trait 25 ('unconventional, eccentric') actually negative, -0.05.

(vii) *Cordial Type*. This type of personality is 'talkative,' 'sociable,' 'imaginative,' and 'emotional'—again a cyclothymic group of traits. Judged by the saturations the underlying factor would seem to correspond roughly with Cattell's factor 5. It is, however, somewhat narrower, since that has saturations almost as large for 'cheerfulness,' 'energy,' and 'frankness,' and is named 'surgency.'

(viii) *Facile Type*. The last type is 'indolent, relaxed, or easy-going,' 'cheerful,' and not 'aesthetically fastidious.' Of the factors in Cattell's list, factor 10 ('Bohemianism *versus* Spiessburger concernedness') perhaps comes nearest to the notion of a cheerful, easy-going type. But the saturations are not quite what we should expect. This doubtless is because his two preceding 'cyclothymic' or extravertive factors (2 and 7) have apparently left very little residual correlation to be accounted for in the relevant traits.

Both the bipolar and the group factor analyses suggest the further presence of several very narrow factors, linking, as a rule, only two or three traits. These appear to be an outcome of the way in which the traits were selected. Each of the factors previously known (we are told) was represented by "a minimum of two traits." It is these duplicated traits that are responsible for some of the largest of the observed correlations—correlations which are so specific that they show little tendency to vanish even when five or six common factors have been extracted. Consider, for example, the following pairs: 3 and 34 ('self-sufficient' and 'self-contained'), 24 and 35 ('emotionally mature' and 'immature'), 18 and 28 ('timid' and 'shy'), 12 and 19 ('soft-hearted' and 'good-natured'), 13 and 26 ('placid' and 'calm, phlegmatic'), 3 and 4 ('attention-getting' and 'assertive'), 18 and 16 ('retiring' and 'silent, introspective'), 15 and 24 ('self-willed' and 'impatient'). Nearly all these pairs yield large correlations, numerically between 0.60 and 0.75. Consequently, however well the main group factors fit the general run of the correlation-table, there are bound to be a few isolated residuals demanding small supplementary factors.

IV. SUMMARY AND CONCLUSIONS

1. The table of correlations between 35 conative traits, factorized by Dr. Cattell in accordance with Thurstone's centroid method, has been refactorized by Burt's method of simple summation to obtain bipolar factors and group factors. The chief differences in procedure are that, where necessary, the traits have been reversed so that their polarity has the same direction throughout, the communalities have been assessed by successive approximation, and the group factor analysis has been based on the bipolar analysis.

2. From a practical standpoint the method here proposed seems at once simpler and speedier. Of the time required for Cattell's calculations scarcely one twentieth was needed for the calculations described in the present paper.

3. From a theoretical standpoint the results obtained are far less equivocal, and, being based almost entirely on automatic calculation, are unaffected by arbitrary or subjective influences. Except for three high correlations due to partial identity of traits, four unrotated factors, instead of twelve, suffice to explain the correlations. It is concluded that the remaining eight bipolar factors obtained by Cattell are either artefacts or else devoid of statistical significance.

4. Both the bipolar and the group factors point to the same general classification of traits or types. There appears to be first a general factor common to all the traits ; secondly, two large bipolar factors cross-classifying traits and personalities into (a) euthymic or dysthymic, and (b) organized or unorganized ; and, thirdly, a series of smaller group factors, leading to eight sub-types, more or less corresponding with those obtained in previous researches. The analysis as a whole thus yields a hierarchical classification of personality-qualities instead of a miscellaneous list of unrelated group factors.

REFERENCES

1. Cattell, R. B. 'The primary personality factors in women compared with those in men.' *Brit. J. Psych. (Stat. Sect.)*, I, 1948, 114-130.
2. Cattell, R. B. 'The description of personality : principles and findings in factor analysis.' *Amer. J. Psych.*, LVIII, 1945, 69-90.
3. Cattell, R. B. *The Description and Measurement of Personality*. New York : World Book Company, 1946.
4. Burt, C. 'The general and specific factors underlying the emotions,' *Brit. Ass. Ann. Rep.*, 1915, 694-696.
5. Burt, C. 'Factor Analysis by Simple Summation : Note on Working Procedures.' University College, Laboratory Notes for Research Students, 1930.
6. Burt, C. 'The factorial analysis of emotional traits.' *Character and Personality*, VII, 1939, 238-254, 285-299.
7. Burt, C. *The Factors of the Mind*. London : University of London Press, 1940.
8. Thurstone, L. L. *The Vectors of Mind*. Chicago : University of Chicago Press, 1935.
9. Thurstone, L. L. *Multiple Factor Analysis*. Chicago : University of Chicago Press, 1947.
10. Thomson, G. 'Cattell's study of personality,' *J. Educ. Psychol.*, XXXVIII, 1947, 273-282.

BOOK REVIEWS

Construction and Analysis of Achievement Tests. By Dorothy C. Adkins and others.
Washington : Government Printing Office, 1947. Pp. xvii + 292. \$1.25.

This book gives a comprehensive account of the scientific development of written tests for personnel selection. It was conceived as a training handbook for new members of the examination staff of the U.S. Civil Service Commission and War Department, and has been used primarily for this purpose. It should be read by everybody in this country who is engaged either in personnel selection or in training students for that work. The language is simple and clear. Relatively few technical terms are used ; and they are defined both in the text and in an unusually full index. Abundant illustrations and examples are given. The book will thus be found very readable by non-psychologists. If it finds its way into the hands of senior Civil Servants and business men responsible for the methods of selecting new entrants to their respective organizations, it should convince them that in the United States scientific selection is no longer an intriguing possibility, but an established fact. Based as they are on wide experience, the recommendations in the book are eminently practical, and take due account of public feeling. For example, it is emphasized that tests must appear relevant and possess 'face-validity,' as well as being valid in fact. "If an item-constructor innocently includes Socrates in a distracter, he should not be surprised if the examination is later publicized as absurdly impractical because it inquires into the competitors' knowledge of Greek philosophy." Again, it is suggested that, when raw scores have been converted into standard scores, an arbitrary number of marks should be added to each, so that no-one is credited with zero.

The emphasis in this book is on achievement tests, which are used far more extensively in the United States than in the United Kingdom ; but most of the points made are of general application. It is assumed throughout the book that limited-response tests are best. The advantages of this type of test, particularly as regards objectivity of scoring, are numerous and important ; but certain types of inventive-response test, e.g., Arithmetic and Abstractions items, can be easily scored and have other advantages. The insistence of psychologists on making appropriate experiments before taking decisions is clearly brought out, as is the fact that measures of job competence should be established in the very early stages of constructing a test. Many promising tests have probably failed to come to anything through the absence of any reliable criterion against which to validate them. The importance is stressed of basing test construction on a thorough job-analysis and of "aiming at a detailed and specific break-down of areas allocated to the written test in the total examination plan." It is pointed out that written selection tests of any kind may be inappropriate if the jobs are too heterogeneous, if the field of work is poorly defined so that proficiency ratings are unreliable, or if the most important relevant factors cannot be assessed by written tests.

Miss Adkins makes it plain that test construction is a highly skilled job, requiring close co-operation between 'test technicians' and 'subject-matter specialists.' The former cannot be expected to have sufficient knowledge of the subject-matter to construct comprehensive and valid tests on their own; neither can the experts in the subject-matter produce technically sound tests without assistance from the test technicians. The right course is for both parties to co-operate at all stages ; and on page 4 she puts forward a detailed plan based upon her own experience of how this co-operation is best achieved. She rightly claims that, with skill in their construction, objective tests need not be limited to testing knowledge of facts, as is sometimes alleged. They can also test ability to use information, appreciation of principles, the extent of insight into the basic reasons for relationships and the power to perform other complex mental processes. In view of the skill and care required in inventing test items, Miss Adkins suggests that a reasonable output is 15-20 'normal' items or four 'complex' items per day ; and I would agree that this is as much as anybody could be expected to produce over a long period. Chapter V presents a discussion of the development of performance tests which will be of special interest to industrial psychologists and managers.

In a book of this size a few minor blemishes are inevitable. The non-statistician will have no fault to find with Chapter III on "Basic Statistical Tools"; but the statistician will complain of certain inaccuracies. For example, leptokurtosis and platykurtosis are wrongly defined and illustrated (p. 103); and a large-sample estimate of the standard error of r should not be used for a correlation of 0.851 with a sample of 42 (p. 133), since, when r is greater than 0.8, the sampling distribution becomes very skew. The Kuder-Richardson theoretical estimates of reliability (pp. 153-154) are interesting and useful; but it is arguable that the reliability of a test under normal working conditions is best measured by the correlation between two equivalent forms of the test given to the same subjects at an interval of (say) a fortnight. In the section on item-analysis (pp. 180-186) it is surprising that there is no mention either of Horst's second method (*Psychometrika*, 1936, pp. 229-245) using r (criterion) divided by r (test), or of the Richardson and Adkins short-cut method (*J. Educ. Psych.*, 1938, pp. 547-552), which are among the most promising. These criticisms, however, are trivial when it is remembered what a wealth of information has been packed into a volume which is still small enough to be used as a practical handbook.

E. A.

Mathematical Theory of Human Relations. By N. Rashevsky. Bloomington (Indiana): Principia Press. Pp. xiv + 202. \$4.

Those students of society who accumulate quantitative facts, are occasionally in need of classificatory ideas: Rashevsky offers such ideas in profusion. For example, he supplies two quantitative definitions of personal freedom. The simpler of them is $(w_0 - w)/w_0$, where w_0 is the maximum work which the individual could perform, and w is the work which in fact he does perform in the same time in order to earn his living.

A section in this book ordinarily begins with a set of assumptions defining an abstract type of society. Some of these assumptions are ingeniously chosen to represent current beliefs; others seem to express Rashevsky's personal intuitions. The logical consequences of the set of assumptions are then deduced by the aid of mathematics in the hope of reaching a formula which can be compared with observable facts. Sometimes the facts are available, and a brief comparison is attempted; sometimes the theory has to be left as a stimulus to the future collector of relevant facts. In either case, Rashevsky passes on to a different set of assumptions defining another type of society; and the game begins anew.

When arranging his chapters, Rashevsky ignored the maxim that one should proceed from the particular to the general, and from the simple to the complicated. He admits this, with disarming candour, in the preface. Thus chapters I and II treat a society arranged as it actually is, not in broad social classes, but in the finest gradation of individual influence. The requisite mathematics is the integral equation; and it appears in chapters I and II in rather elaborate forms. Fortunately for those readers who are not familiar with integral equations, the argument of the book is not continuous: several fresh starts are made. Chapters I and II may quite well be omitted. Those who wish to pick out an exceptionally easy sample of Rashevsky's theorizing may do well to begin with chapter X on urban and rural population, or with chapter XVII on density of population, or with chapter XIX on individualism and collectivism; for these three chapters begin with elementary algebra. In general the mathematics in the book is as accurate, clear, and neat, as one could wish; but there are a few misprints and a few ugly notations.

In most theories, in any part of science, some compromise has to be made between truth and tractability. The compromise usually preferred by Rashevsky, from chapter III onwards, is to imagine a society composed of two or three classes, all the individuals in a class being alike in some specified characteristic. Let x, y, z , be the numbers of persons in the several classes. The total population N is $x + y + z$; and may be constant or variable. The further properties of this fictitious society are specified by a differential equation giving dx/dt in terms of x, y , and z , and by similar equations giving dy/dt and dz/dt . Such a treatment by distinct classes appears to this reviewer as a reasonable compromise; for he has himself independently used it in a discussion of war-moods (*Psychometrika*, XIII, 1948).

A representation of all the observed gradations of individuals in some specified quality could be achieved by allowing the number of classes to increase indefinitely. The set of simultaneous differential equations would then pass over into an integro-differential equation of the sort already mentioned by Rashevsky in chapters I and II, which it is desired to avoid.

The particular applications which Rashevsky makes of this general scheme are very varied. The classes are sometimes specified by their influence, sometimes by their wealth, sometimes as military, industrial, agricultural or artistic, sometimes by their fertility, sometimes as belonging to different nations; but the leading idea, which runs through most of the specifications, is that society is divided into a small class of 'actives' who exert their influence over a large class of 'passives.'

Psychologists may wish to excuse themselves from attending to some of these chapters by calling them economics. I would suggest, however, that the boundary between social psychology and economics is about as artificial as the boundary between chemistry and physics, in the sense that practical problems about society usually involve both economics and psychology, just as practical problems about industrial materials usually involve both chemistry and physics. The academic etiquette of not poaching on a colleague-professor's preserves does not deter Rashevsky from discussing mixed problems. Some of his more psychological themes are: influence and fashion in chapters I to IV, envy or jealousy in chapter VII, freedom in chapter IX, class-solidarity in chapter XIV, selfish or public-spirited motives in chapters XIX and XX, migration in chapter XXI, and conflict in chapters XXII and XXIII.

No one of the arguments is brief enough to be quoted here in full; but three samples will be summarized.

First, a sample from chapter XIX about selfishness and public spirit. Let there be a closed society consisting of two individuals, presumably the sole inhabitants of an island. Let them perform amounts of work, y_1 and y_2 , which they dislike, and thereby produce jointly an amount X of goods, which they desire. It is assumed that

$$X = a_1 y_1 + a_2 y_2 + a_3 y_1 y_2,$$

where a_1, a_2, a_3 are constants. The term $a_3 y_1 y_2$ represents the effect of co-operation. The net satisfactions, S_1 and S_2 , which they respectively derive from the whole affair, are assumed to be

$$S_1 = A_1 \log(C_1 X) - B_1 y_1, \quad S_2 = A_2 \log(C_2 X) - B_2 y_2,$$

where $A_1, B_1, C_1, A_2, B_2, C_2$ are constants. It is then proved that if each person tries to maximize his own satisfaction S_1 , or S_2 , they produce less goods than if both try to maximize their joint satisfaction $S_1 + S_2$.

Next, a sample about class-solidarity, heredity, and revolution. At an initial time, $t = 0$, a nation is supposed to consist of an active minority x successfully controlling a passive majority y . Some of the children of each class have innately the mental characteristics of the other class. But, if class-solidarity is a rigid custom, these aberrant children will nevertheless be retained in the class to which they were born. Then a time may come at which a diluted class of actives may no longer be able to control a fortified class of passives; so that a revolution occurs. This theme, imbedded among many variants and elaborations, occurs in chapters XIV, XVII, and XVIII. The rate of increase of x is specified sometimes simply by $dx/dt = Ax$ and sometimes by an equation which has been much used by biologists, namely

$$dx/dt = Ax - Bx^2,$$

where A and B are constants. The changes in y and in the numbers of aberrant offspring are similarly specified.

Thirdly, a sample about war. In chapter XXII Rashevsky offers a theory of a war which goes on by destruction in accordance with the equations

$$dM/dt = -k_2^2 N, \quad dN/dt = -k_1^2 M,$$

in which M and N are the opposed populations, and k_1^2, k_2^2 are constants. As two psychologists have told me that these equations resemble my theory of arms-races (1939, *Brit. J. Psych. Mon. Sup.*, No. 23, or its revision in a 1947 microfilm), I should point out that during an arms-race the constants have the other sign, and M and N have other meanings.

Rashevsky supposes that the war comes to an end when $N = 0$, that is, when one side is exterminated. Such ends have been rare. Among 97 considerable fatal quarrels I find only one that ended by extermination: that was the massacre of the Janissaries in Turkey in A.D. 1826. Rashevsky, however, soon modifies his hypotheses into the supposition that the leaders of one side were gradually killed, until those of them who survived were too few to control the 'passive' members of their population. The war in Sumatra, from 1873 till about 1908, between the Dutch and the ancient sultanate of Achin, was of that type; but the type has been unusual, and in particular is quite unlike the two World Wars.

If all the towns in a country are ranked in order of their population, the largest being placed first, then the product of each town's rank and population is roughly the same for all. According to Lotka (*Elements of Physical Biology*) this remarkable empirical law was published by F. Auerbach in 1923. Rashevsky, on the contrary, attributes it to Zipf in 1941. Rashevsky attempts several different explanations; although they run to about 80 equations, they do not conspicuously succeed. Thus the observed fact is left more mysterious than it was before.

Of the 17 diagrams in the book, 12 exhibit observed facts. Rashevsky's comparisons between fact and theory are, however, not impressive: they often establish a *prima facie* case for further enquiry, but seldom or never a statistical proof. Indeed, he repeatedly disclaims any intention of dealing critically with the facts; and he restricts his purpose to pointing out the interesting consequences which follow logically from specified assumptions. Nevertheless, six of his diagrams could easily have been drawn better; for their abscissæ are spacings of nations in an arbitrary order, and they have two ordinates: one for theory, one for fact. This type of comparison had better be presented as the usual correlation-diagram, free from any arbitrary spacing. Rashevsky does not mention the significance of correlation coefficients. On the other hand, he gives on pp. 145-7 a highly suggestive empirical comparison of the international influence of six countries by way of an index-number, defined to be $(\text{population})^3 (\text{area})^{-1}$.

In short, the words that Sir Oliver Lodge applied to the *Philosophical Magazine*, which he edited, may also be applied to Rashevsky's book: "all of it interesting, and some of it true." The author, in a long preface, has reviewed his own work in a spirit so candid and critical that we may expect many future improvements from him.

L. F. R.

Handbook of Tests for Use in Schools (Second Edition). By Sir Cyril Burt. London: Staples Press. Pp. xvi+110. 10s. 6d.

This is the fourth impression of Burt's well-known book of standardized tests for assessing the intelligence and educational attainments of school children. A valuable addition to the new edition is the list of tests for the Terman-Merrill Revision of the Binet Scale: the list shows the new age-assignments for English children, obtained by the investigations of the Binet Committee.

As is rather obvious from the type and paper, the book has been "made and printed in Holland." But teachers, educational psychologists, and school medical officers will be glad to have once again so useful a volume, which unfortunately has been out of print since the beginning of the war.

E. K.